

ON THE CONDITIONAL APPROACH TO SELECTIVE INFERENCE

A thesis presented for the degree of
Doctor of Philosophy
in the
Department of Mathematics
of
Imperial College London
by

Daniel García Rasines

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Selective inference methods have garnered considerable attention in the past few years. One of them, the so-called conditional approach, under which inferential statements are derived from a conditional model of the data, provides a particularly simple and widely applicable solution to the problem. The aim of this thesis is to carry out a detailed analysis of this approach from several angles. We analyse it conceptually from frequentist and Bayesian viewpoints, establish some theoretical results, and propose routes for numerical implementation of the approach in complex regimes. Two key components of this work are the derivation of a class of approximate probability-matching priors for selection models, and the analysis of a simple alternative to data splitting in regression models which produces more powerful inferences than the latter method. We provide an empirical validation of all the methodological proposals.

Acknowledgements

Firstly, I would like to thank my supervisor, Alastair Young, for his guidance during this PhD. His regular and insightful feedback contributed greatly to the development of this work, as did his constant support and positivity. This was a challenging project, and I would not have completed it without his excellent supervision.

I thank the Department of Mathematics of Imperial College London for generously sponsoring this PhD, as well as my attendance to the *4th WHOA-PSI* conference.

I thank my family and friends for their constant support, so necessary, especially in the difficult times.

Lastly, I thank the members of the IC Cross Country and Athletics Club, Thames Valley Harriers, and the IC Choir for making my stay in London so enjoyable.

Daniel García Rasines

Contents

1	Introduction	11
1.1	Selection models	11
1.2	The conditional approach	14
1.3	Randomised selection	15
2	Conceptual aspects	18
2.1	Sampling models	18
2.1.1	Ancillarity in selection models with nuisance parameters	22
2.2	Bayesian inference	25
2.2.1	Fixed and random parameters	29
3	Some results about selection models	32
3.1	Point estimation	32
3.2	Asymptotic selective normality	38
3.3	Appendix	44
3.3.1	Proof of Lemma 1	44
3.3.2	Proof of Theorem 2	46
4	Non-informative priors for selected parameters	51
4.1	Probability matching in normal models	51
4.2	One-dimensional exponential family	57
4.3	Non-informative priors in the presence of nuisance parameters	62
4.3.1	Inference for a selected normal mean with an unknown variance	63

4.3.2	Inference for the winner	64
4.4	Appendix	66
4.4.1	Proof of Proposition 3	66
4.4.2	Derivation of the probability-matching prior with data splitting	67
5	Splitting methods for regression models	68
5.1	Introduction	68
5.2	Inferential targets	70
5.3	Splitting methods	72
5.4	Theoretical comparison	73
5.5	Generalised (U, V) decomposition	75
5.6	Simulation study	77
5.6.1	Selection power	79
5.6.2	Selection stability	79
5.6.3	Inference for selected coefficients	80
5.6.4	Inference for projection parameters	81
5.7	Appendix	83
5.7.1	Simulation results	83
5.7.2	Proof of Proposition 4	90
5.7.3	Proof of Theorem 3	90
6	Numerical implementation	94
6.1	An Importance Sampling algorithm	94
6.2	Examples	98
7	Discussion and future work	101

List of Figures

1.1	Evolution of the Fisher information of the selective univariate model as γ varies from 0 to 3 for $\sigma^2 = 10^{-2}$ and $\theta = -1, 0, 1$. The solid line is the total information and the dashed one the information contained in V only.	17
2.1	Coverage of credible intervals under repeated sampling from fixed values of θ derived from the unadjusted posterior (the red line corresponds to sampling with selection and the blue one without it), and from the selective posterior (green).	28
3.1	MLE (black) and MOE (blue) as a function of y for one and two-sided selection, with $n = 10$, $t = 2n_1^{-1/2}$, and $\gamma = 0.25, 0.5, 0.75$	37
3.2	In solid black, distribution function of $\hat{u}$ for different values of γ and n . In dashed black, the distribution function of the approximation provided by the theorem.	43
4.1	Left: uniform prior (black), probability-matching prior for $y = 0.2$ (red), and Jeffreys prior (blue) for the normal model. Right: the resulting posteriors for $y = 0.2$ (the red density overlaps the blue one in this panel).	56
4.2	Left: non-selective Jeffreys prior (black), probability-matching prior for $y_1 = 4$ and $y_2 = 1$ (red), and Jeffreys prior (blue) for the binomial model. Right: the resulting posteriors (the red lines overlap the blue ones).	58
4.3	Example 4.2.1. Coverage of the interval $(-\infty, \Pi^{-1}(\alpha Y_1, \dots, Y_n)]$ as a function of α for the non-selective Jeffreys prior (red), the probability-matching prior (orange), and the selective Jeffreys prior (green). The orange lines partially overlap the green ones.	60
4.4	Example 4.2.2. Coverage of the interval $(-\infty, \Pi^{-1}(\alpha Y_1, \dots, Y_n)]$ as a function of α for the non-selective Jeffreys prior (red), the probability-matching prior (orange), and the selective Jeffreys prior (green). The orange lines partially overlap the green ones.	61
4.5	Estimated empirical CDFs of $\Pi(\mu_0 Y_1, \dots, Y_n)$ for the prior $\pi(\mu, \sigma^2) \propto \sigma^{-1}$ (red) and for Jeffreys prior (green).	64
5.1	Power of the knockoff applied to the complete dataset (black), a data split (red), and a randomised version of the data (blue). Titles indicate (f, ρ, p)	85

5.2	Power of stability selection applied to the complete dataset (black), a data split (red), and a randomised version of the data (blue). Titles indicate (f, ρ, p)	86
6.1	In black, $L_S(\theta; y)/L_S(1; y)$. In red, Monte Carlo approximation. In blue, non-selective version.	99
6.2	Empirical posterior distribution function of ψ obtained from Metropolis-Hastings sampling using: the correct acceptance probabilities (black), and approximations of the acceptance probabilities derived from the suggested method (red). The non-selective posterior distribution function is plotted in blue.	100

List of Tables

4.1	Estimated coverages of $(-\infty, \Pi^{-1}(\alpha Y)]$ for the normal-location model derived from the uniform (U) and Jeffreys (J) priors.	56
4.2	Coverage and average length of 0.9-credible intervals for the mean giving the largest sample, derived from likelihoods $L_1(\theta)$ and $L_2(\theta)$ and Jeffreys prior.	65
5.1	True positive rate of the selection algorithms applied after data splitting (DS) and randomisation (R), normalised by the true positive rate of selection applied to the full dataset.	83
5.2	Means and standard deviations of the estimated selection probabilities for fixed values of (Y, X)	84
5.3	Coverages of confidence intervals for the selected coefficients; knockoff selection.	87
5.4	Coverages of confidence intervals for the selected coefficients; stability selection.	87
5.5	Average length of confidence intervals for the selected coefficients; knockoff selection.	88
5.6	Average length of confidence intervals for the selected coefficients; stability selection.	88
5.7	Coverages of confidence intervals for the coefficients of the projection parameters; stability selection.	89
5.8	Average length of confidence intervals for the coefficients of the projection parameters; stability selection.	89

Chapter 1

Introduction

Classical statistical methods are designed to provide error guarantees in situations where the objectives of the analysis are specified before collecting the data. In many contemporary problems, however, such idealised settings are the exception rather than the norm. A more realistic situation is one in which the statistician performs an exploratory analysis of the data before selecting a set of relevant inferential questions to address. It is well known that failing to acknowledge this adaptivity in the subsequent inferences yields the reported error assessments invalid. For example, type I error guarantees of testing procedures are lost. This problem, commonly known as selection bias, has received considerable attention in the past few years, especially in the regression literature. There are multiple angles from which one can approach it. In this thesis we analyse in detail one of them, the conditional approach, which we introduce and revise in this chapter.

1.1 Selection models

A standard problem involving selection is the so-called selected mean problem. Suppose that $Y_1, \dots, Y_n$ are independent normal observations with different unknown means $\theta_1, \dots, \theta_n$ and known common variance, and that we seek inference for the means of the largest k observations for some $k < n$. A problem like this can arise, for example, in a clinical trial experiment, where a set of n treatments are tested and only the most promising ones are put forward for further investigation. In this situation, it is clear that reporting standard estimates and confidence intervals for the selected parameters ignoring selection, what is commonly known as a face-value approach, will likely lead to flawed conclusions, as the largest observations tend to overestimate their corresponding means. This particular case of selection bias is known in the literature as winner's curse.

More generally, suppose that we have data $Y \in \mathcal{Y}$, whose sampling distribution we model by some parametric family $\{F(y; \theta) : \theta \in \Theta\}$, where $F(y; \theta)$ is the distribution function of Y under θ . We assume throughout that $\mathcal{Y}$ and Θ are both subsets of a Euclidean space, and that all the distributions in the family are dominated by the Lebesgue or counting measures. The corresponding Lebesgue densities or probability mass functions will be denoted by $f(y; \theta)$. In addition, we assume that there exists a set of m potential parameters of interest, $\{\psi_1(\theta), \dots, \psi_m(\theta)\}$, from which at most one is to be selected for inference after observing the data, possibly by an artificially randomised procedure. For example, in the previous problem, the set of potential parameters of interest would contain all the subsets of $\{\theta_1, \dots, \theta_n\}$ of size k . Thus, we assume that there exist functions $p_i : \mathcal{Y} \rightarrow [0, 1]$, $i = 1, \dots, m$, such that, having observed $Y = y$, $\psi_i(\theta)$ is selected for inference with probability $p_i(y)$.

In this formulation of the problem, at most one of the parameters can be selected, so we will generally drop the subscript and write the selected parameter under consideration simply as $\psi \equiv \psi_i(\theta)$, and the corresponding selection probability as $p(y) \equiv p_i(y)$. The latter function, which encodes all the relevant information regarding the selection mechanism of ψ , will be referred to as the selection function. The following examples illustrate the framework.

Example 1.1.1. File drawer problem. Let $Y_1, \dots, Y_{n_1} \in \mathbb{R}$ be a random sample from a distribution with scalar parameter θ , and suppose that we use this sample to test $H_0 : \theta = 0$ against $H_a : \theta > 0$, because we are only interested in θ if it is positive. If the test results in rejection, we collect a second set of samples $Y_{n_1+1}, \dots, Y_n$ to obtain more information about θ . At the end, the whole data vector is $Y = (Y_1, \dots, Y_n)^T$. If, for example, the testing procedure rejects H_0 whenever $S(Y_1, \dots, Y_{n_1}) > t$ for some statistic S and some threshold t , the selection function is $p(y) = \mathbf{1}\{S(y_1, \dots, y_{n_1}) > t\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

Now, suppose that for a certain population under study the true parameter is $\theta = 0$, and that a process like the one described before is carried out by several independent researchers. If only those analyses for which the initial test rejected the null are published, perhaps by journal requirements, the literature concerning this parameter will be filled with results that overstate the true effect size, leading to the false belief that θ is positive, when in reality it is not. This is a stylised example of what [Rosenthal \(1979\)](#) termed the *file drawer effect*, also known as *publication bias*, which occurs when the decision of making a scientific finding public is influenced by the outcome of the analysis.

Example 1.1.2. Regression model. Let $Y \in \mathbb{R}^n$ be a response vector and $X \in \mathbb{R}^{n \times p}$ be a known, fixed design matrix containing the observed values of p covariates. Suppose that a variable-selection algorithm is used to select a non-empty subset $s \subseteq \{1, \dots, p\}$ of the

covariates, and that we wish to provide inference for the best linear predictor of Y given only the selected covariates, $\psi = \{X(s)^T X(s)\}^{-1} X(s)^T \mathbb{E}_\theta[Y]$, where $X(s)$ is the submatrix of X containing the observations of the selected covariates. Denoting by $E(s) \subseteq \mathbb{R}^n$ the set of observations for which the variable-selection algorithm would have produced the same set of selected covariates, we can write the selection function as $p(y) = \mathbf{1}\{y \in E(s)\}$. Inference for these models is the central topic of Chapter 5.

Like many other fundamental problems in statistics, selection issues were first studied by [Fisher \(1934a\)](#), who considered the sampling bias arising in albinism studies. The type of problem considered by Fisher is one in which independent samples $Y_1, Y_2, \dots$ from a given population are generated and each sample is recorded or not with probability $w(y_i)$, so that the data available for inference constitutes a random sample from a weighted distribution $g(y; \theta) \propto f(y; \theta)w(y)$. There is a large body of literature dedicated to the study of these models. A good survey is given by [Rao \(1985\)](#).

Although these problems can be considered under the previous framework, by defining the selection function $p(y_1, \dots, y_n) = \prod_{i=1}^n w(y_i)$, they are different in nature to the kind of problems studied in this work. Conceptually, the main difference is that in our problems selection refers to the decision of providing or not inference for a given parameter, while in the other case the inferential objective is decided in advance, and selection refers to a bias of the sampling mechanism. As we will see in Section 2.2, a classical Bayesian analysis would treat these problems differently, adjusting the inferences in the latter case but not in the former. Mathematically, the selection functions considered here cannot usually be broken up into marginal selection functions for the different samples. As in Example 1.1.1, the decision of providing inference for a given parameter will depend on some summary statistic of the full dataset or a subset of it. The implications of this are important. When $p(y_1, \dots, y_n) = \prod_{i=1}^n w(y_i)$ and $Y_1, \dots, Y_n$ are independent, the distribution induced by the selection rule can be factorised into independent components. By contrast, the selection rules considered here induce non-trivial dependencies between the observations, and such dependencies do not vanish when the sample size increases except in trivial circumstances.

In this work we will only consider situations where the selection function of the selected parameter is known, at least to a high degree of precision. Inference with a fully unknown selection function is ill-posed, as the model becomes highly non-identifiable. This problem has been considered by some authors (e.g. [Bayarri and DeGroot, 1987](#)) in cases where the selection function is assumed to be of the form $p(y) = \prod_{i=1}^n w(y_i)$.

1.2 The conditional approach

Different approaches to statistical inference lead to two opposing views as to the correct analysis of the data in the presence of selection. On the one hand, frequentist methods evaluate the accuracy of inferential procedures with respect to the sampling distribution of the data at a fixed value of the parameter. Since selection modifies the sampling distribution by favouring data points with higher selection probability, it is clear that the reported accuracy should be appropriately modified. On the other hand, Bayesians typically adopt the view that, once the data has been observed, the recognition that a different realisation could have resulted in a different inferential problem, or in no problem at all, should have no effect on the inference (Dawid, 1994). The latter position, however, has recently been challenged by some authors (Yekutieli, 2012; Harville, 2021). We analyse this discrepancy in Section 2.2.

Adoption of the first viewpoint leads to adherence to the so-called conditional approach, which advocates that inference for the selected ψ should be based on the conditional distribution of the data given selection. Such distribution has density or mass function

$$f_S(y; \theta) = \frac{f(y; \theta)p(y)}{\varphi(\theta)}, \quad \varphi(\theta) = \mathbb{E}_\theta[p(Y)]; \quad (1.1)$$

where the normalising constant $\varphi(\theta)$ is the probability that ψ gets selected when θ is the true parameter. The motivation for basing inference on (1.1) is that, under repeated sampling from a given $F(\cdot; \theta)$, if inferences are only provided for those samples that get selected, the reported error assessments are well-calibrated. For example, nominal $1 - \alpha$ confidence sets would contain the true parameter at least $(1 - \alpha)\%$ of the times they are reported.

The conditional approach has received much attention in recent years, particularly in the context of inference after variable selection (Example 1.1.2). Interest on this approach increased after the realisation that some widely used selection rules produce simple selection events, like affine sets, which can be handled with relative ease. This line of work began with Lockhart et al. (2014) and has been explored by numerous authors, including Lee and Taylor (2014); Loftus and Taylor (2014); Lee et al. (2016); Fithian et al. (2017); Tibshirani et al. (2018); and Panigrahi et al. (2020).

In the remainder of this work, we will refer to the conditional distribution of Y given selection as the selective distribution, and to the corresponding likelihood function, $L_S(\theta) = f_S(y; \theta)$, as the selective likelihood. When writing probabilities, expectations, etc., we will

denote the selection event by S . So, for example, for continuous data, we would write

$$\begin{aligned}\mathbb{E}_\theta[Y] &= \int_{\mathcal{Y}} y f(y; \theta) dy; \\ \mathbb{E}_\theta[Y|S] &= \int_{\mathcal{Y}} y f_S(y; \theta) dy.\end{aligned}$$

An intuitive way to interpret the conditional approach is as a form of information splitting. For a given ψ , let R be the random variable that takes the value 1 if ψ gets selected and 0 otherwise. That is, $R|Y \sim \text{Bernoulli}\{p(Y)\}$. Following [Fithian et al. \(2017\)](#), the data-generating process of Y may be thought of as consisting of two stages. In the first stage, the value of R , r say, is sampled from its marginal distribution, and in the second, Y is sampled from the conditional distribution $Y | r$. Since it is R that determines whether we are going to provide inference for ψ or not, inference based on information revealed in stage two is necessarily free of any selection bias, as it removes the information about the parameter provided by R .

[Kuffner and Young \(2018\)](#) argued that the conditional approach is in fact a modern re-expression of well-established Fisherian arguments. Specifically, of what they term the *Fisherian proposition*, based on [Fisher \(1925, 1934b\)](#), which says that relevance of the inference to the observed data is achieved by conditioning on certain sample statistics. Typically, these statistics are required to be uninformative about the parameter of interest, though it is generally accepted that this condition can be relaxed in some circumstances ([Barndorff-Nielsen and Cox, 1994](#), Chapter 2).

1.3 Randomised selection

In the conditional approach, the information provided by the data is divided in two portions, one of which is used for selection (R) and the other one for inference ($Y|R$). There is therefore a trade-off between the power of the selection mechanism (this is, its ability to identify a parameter when it is truly interesting, like a significant effect), and the power of the subsequent inferential method. As we shall see in the following chapters, failing to control the amount of information used in the selection stage can result in inferences with too little power, especially when the true parameter has a low selection probability. In some situations this may not be a concerning issue. For example, if selection is the main goal of the analysis and inference is secondary, or if we are able to obtain an extra set of samples after selection. If, however, powerful inference is required and obtaining new data is infeasible, we need to utilise the available information efficiently. When the data consists of independent and identically distributed (IID) observations, the most natural

way of achieving this is by using only a subset of the observations in the selection stage, and incorporating the remaining ones in the inferential analysis. This is, in fact, a specific case of a very flexible mechanism known as randomised selection, which was suggested by [Tian and Taylor \(2018\)](#).

Randomisation allows us to limit the amount of information used in the initial stage by applying the selection mechanism to a noisy version the data. Formally, we would generate a random variable W , with known distribution and independent of the data, and base selection on $U = u(Y, W)$ rather than Y , where u is some function of the data and the noise, such as $Y + W$. Note that, if $p_U(u)$ denotes the selection function in terms of U , the selection function of the data Y would be computed as $p(y) = \mathbb{E}[p_U(y, W)]$. For IID samples $Y = (Y_1, \dots, Y_n)^T$, the natural option of using only a subset of the samples can be written this way by letting W be a random subset of $\{1, \dots, n\}$ of a given size and letting U be the set of observations with indices in W . Randomised selection will be explored more closely in Chapter 5.

An interesting example of randomisation can be produced when the data is normal. Suppose that $Y \sim N(\theta, \Sigma_Y)$, where $\theta \in \mathbb{R}^n$ is unknown and $\Sigma_Y \in \mathbb{R}^{n \times n}$ is a known positive definite matrix, and suppose that a parameter of interest ψ gets selected if and only if $Y + W \in E \subseteq \mathbb{R}^n$, where $W \sim N(0_n, \Sigma_W)$ is a noise vector independent of Y , 0_n is an n -vector of zeroes, and $\Sigma_W \in \mathbb{R}^{n \times n}$ is a positive definite matrix chosen by the statistician. It is useful to rewrite the data as $Y = (I_n + \Sigma_Y \Sigma_W^{-1})^{-1}(\Sigma_Y \Sigma_W^{-1}U + V)$, where I_n is the $n \times n$ identity matrix, $U = Y + W \sim N(\theta, \Sigma_Y + \Sigma_W)$, and $V = Y - \Sigma_Y \Sigma_W^{-1}W \sim N(\theta, \Sigma_Y + \Sigma_Y \Sigma_W^{-1} \Sigma_Y)$. The reason for doing so is that U and V are independent and selection is defined only in terms of U . Independence follows trivially from the fact that the covariance of U and V is null and that U and V are jointly normally distributed, because (U, V) is a linear transformation of (Y, W) . Thus, the distribution of Y given $U \in E$ is a linear combination of two independent random vectors, one that follows a truncated normal distribution and another one that follows an untruncated normal distribution, and we can study the selective distribution of Y by studying both components separately. One advantage of this is that, in settings where working with the full conditional model $Y \mid \{U \in E\}$ is complicated, we can always use V alone, which follows an easy tractable distribution, to provide inference for the selected parameter. The cost incurred in doing so will depend on how informative the distribution $U \mid \{U \in E\}$ is about ψ .

This decomposition is best understood by considering its one-dimensional version. Suppose that $Y \sim N(\theta, \sigma^2)$, $W \sim N(0, \gamma\sigma^2)$, and $E = [0, \infty)$, for some known constants $\sigma^2, \gamma > 0$. In this case $U \sim N(\theta, \{1 + \gamma\}\sigma^2)$ and $V \sim N(\theta, \{1 + \gamma^{-1}\}\sigma^2)$. The parameter γ has the role of balancing how much information about θ we have in the selection and inferential stages, respectively. As γ approaches 0, the variance of U converges to that of Y , and thus

we are using almost all information to decide whether θ is of interest or not. Conversely, if γ is very large, U contains virtually no information about θ and selection can be imprecise. This behaviour is inverted in the inferential stage; the larger γ is, the more information about θ is contained in the conditional model. We can justify this more rigorously by looking at the Fisher information about θ in the model $Y \mid \{U > 0\}$, given by

$$I_{Y|U>0}(\theta) = I_{U|U>0}(\theta) + I_V(\theta).$$

The Fisher information of V is trivially given by $\{1 + \gamma^{-1}\}^{-1}\sigma^{-2}$. For U , a direct calculation gives

$$I_{U|U>0}(\theta) = \frac{1}{(1 + \gamma)\sigma^2} \left[1 + h_2 \left\{ \frac{\theta}{\sigma(1 + \gamma)^{1/2}} \right\} \right],$$

where $h_2(x)$ is the second derivative of $\log \Phi(x)$ and $\Phi(x)$ is the standard normal distribution function.

Figure 1.1 shows the behaviour of $I_{Y|U>0}(\theta)$ and $I_V(\theta)$ as γ increases in three different settings. All of them have $\sigma^2 = 10^{-2}$. In the first one ($\theta = -1$) the selection probability under the true parameter is very small. We can observe that the information about θ contained in the truncated distribution of U is negligible compared to that contained in V , and therefore basing inference only on V instead of the full data carries virtually no loss of information. In the second setting ($\theta = 0$) the selection probability is equal to 0.5, and we see that there is a substantial loss of information if the value of U is ignored. This is also the case in the last setting considered ($\theta = 1$). In this case the selection probability at the true parameter is almost 1, and we can see that for any value of γ the available information about θ after conditioning on selection is virtually the same as we would have in the absence of selection (σ^{-2}).

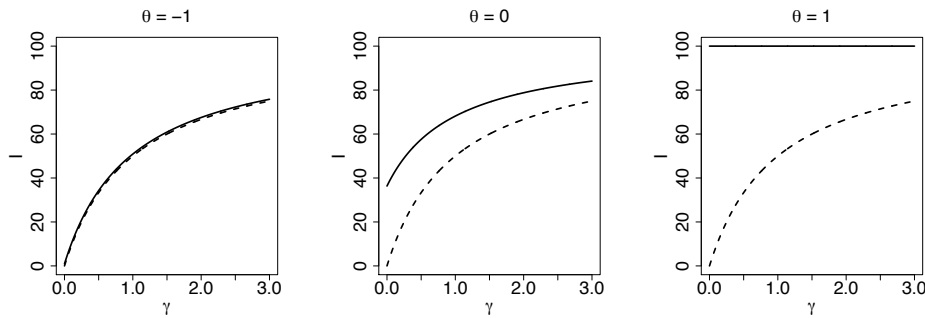


Figure 1.1: Evolution of the Fisher information of the selective univariate model as γ varies from 0 to 3 for $\sigma^2 = 10^{-2}$ and $\theta = -1, 0, 1$. The solid line is the total information and the dashed one the information contained in V only.

Chapter 2

Conceptual aspects

In the first section of this chapter we consider the problem of correct specification of the inferential distribution for selective inference. We show that ancillarity, in the classical sense, is not preserved after conditioning on the selection event. However, we argue that in many cases there are still persuasive reasons for conditioning on ancillaries of the unconditional model: either by reference to how the selection mechanism actually works, or by adopting specific definitions of ancillarity. In the second section we analyse the discrepancy between the classical Bayesian stance on selection and the alternative postures.

2.1 Sampling models

As presented, the conditional approach may appear to be straightforward to apply from a conceptual perspective; if the selection function $p(y)$ is known and one is comfortable with a particular sampling model for the data, then the selective distribution (1.1) is unequivocally determined. However, in this section we argue that, just as the determination of an appropriate sampling distribution in classical statistics is subject to some degree of arbitrariness, so is, in some circumstances, the appropriate choice of the conditional distribution for selective inference. To motivate the discussion, let us consider the following example.

Example 2.1.1. *Random sample size.* Let $Y_1, \dots, Y_{n_1}$ be a random sample from a $N(\theta, 1)$ distribution, where the sample size $n_1 \geq 1$ has been generated from some known distribution $f_{N_1}(n_1)$. Suppose that we use this sample to test whether θ is greater than zero. If we observe $y_1 + \dots + y_{n_1} > 1.96n_1^{1/2}$, say, we collect a second set of observations of fixed size n_2 , $Y_{n_1+1}, \dots, Y_n$, where $n = n_1 + n_2$, which we use to obtain more information about θ . In the joint model for $(N_1, Y_1, \dots, Y_n)$, with selection function

$p(n_1, y_1, \dots, y_n) = \mathbf{1}(y_1 + \dots + y_{n_1} > 1.96n_1^{1/2})$, the conditional distribution of the full data given selection is

$$f_S(n_1, y_1, \dots, y_n; \theta) = \frac{f_{N_1}(n_1)f(y_1, \dots, y_n \mid n_1; \theta)p(n_1, y_1, \dots, y_n)}{\sum_{\tilde{n}_1=1}^{\infty} f_{N_1}(\tilde{n}_1)\Phi(\theta\tilde{n}_1^{1/2} - 1.96)}.$$

Note that in this model the sample size is not independent of θ , as its probability mass function is given by

$$f_S(n_1; \theta) = \frac{f_{N_1}(n_1)\Phi(\theta n_1^{1/2} - 1.96)}{\sum_{\tilde{n}_1=1}^{\infty} f_{N_1}(\tilde{n}_1)\Phi(\theta\tilde{n}_1^{1/2} - 1.96)}.$$

The intuitive explanation of this dependence is that, if the true θ is smaller than zero, we have a higher chance of falsely concluding that $\theta > 0$ if we have a less informative sample (i.e. if n_1 is small), and vice versa. Therefore, inference based on this sampling model would make use of the sample size distribution, and would arrive to a different conclusion had the sample size been decided in advance by the statistician rather than random, which appears counter-intuitive. Instead, it is more reasonable to work with the model

$$f_S(n_1, y_1, \dots, y_n; \theta) = \frac{f_{N_1}(n_1)f(y_1, \dots, y_n \mid n_1; \theta)p(n_1, y_1, \dots, y_n)}{\Phi(\theta n_1^{1/2} - 1.96)}.$$

This model conditions on selection *after* the sample size has been observed, which therefore remains independent of the parameter after conditioning. This reflects more closely the selection process, since the selection test was designed to achieve a certain significance level conditionally on the observed n_1 . That is, the information about θ in the selection step is interpreted according to the conditional model $Y_1, \dots, Y_{N_1} \mid N_1$, rather than to the unconditional one.

Example 2.1.1 illustrates a key issue about selection models: conditioning on selection can create artificial dependencies in the data that do not reflect the nature of the sampling process. Formally, a statistic A which is distribution constant unconditionally will typically depend on the parameter after conditioning on selection. However, it is clear that whether a sample is selected or not cannot make A informative. To avoid this, we have to ensure that the selective distribution reflects the selection process as closely as possible.

Suppose that, unconditionally, the data Y can be reduced to a minimal sufficient statistic (T, A) , where A is distribution constant. In such cases A is said to be ancillary for θ , and the Conditionality Principle, a formal re-expression of the Fisherian proposition, asserts that the information provided by Y about θ is equivalent to that provided by T given the observed value of A (Birnbbaum, 1962). If the selection process complies with this principle, so that the information about θ in the selection stage is interpreted via the conditional

distribution $T|A$, as in Example 2.1.1, then it is more reasonable to define the selective distribution by

$$f_S(t, a; \theta) = \frac{f_A(a) f_{T|A}(t|a; \theta) p(t, a)}{\varphi(\theta; a)}$$

rather than

$$f_S(t, a; \theta) = \frac{f_A(a) f_{T|A}(t|a; \theta) p(t, a)}{\varphi(\theta)},$$

where $p(t, a) = \mathbb{E}[p(Y) | t, a]$ denotes the selection function in terms of $(T, A) = (t, a)$, which independent of θ by sufficiency, $f_A(a)$ and $f_{T|A}(t | a; \theta)$ are the densities or probability mass functions of A and $T | A$, and $\varphi(\theta; a) = \mathbb{E}_\theta[p(T, A) | a]$ is the selection probability under θ given $A = a$. Ancillary statistics arise, for example, in models with an underlying group structure, such as location models.

Example 2.1.2. *Location model.* Let $Y_1, \dots, Y_n$ be a random sample from a location model with density $f(y_i; \theta) = g(y_i - \theta)$, $\theta \in \mathbb{R}$. In this model, the Conditionality Principle dictates that sample evidence about θ should be interpreted via the conditional distribution of $T = \hat{\theta}$ given $A = (Y_1 - \hat{\theta}, \dots, Y_n - \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator of θ and A is the configuration statistic. The density of $\hat{\theta} | a$ admits the simple expression

$$f_{T|A}(t | a; \theta) = c(\theta, a) \prod_{i=1}^n g(t + a_i - \theta),$$

where $c(\theta, a)$ is a normalising constant. Suppose that θ is selected for inference if the p -value

$$u(T, a) = c(0, a) \int_{\hat{\theta}}^{\infty} \prod_{i=1}^n g(\tilde{t} + a_i) d\tilde{t}$$

is below some level α . Then, the selective density of $(\hat{\theta}, A)$ should be

$$f_S(t, a; \theta) = \frac{f(a) c(\theta, a) \prod_{i=1}^n g(t + a_i - \theta) p(t, a)}{\mathbb{P}_\theta(u(T, a) \leq \alpha | a)},$$

where $p(t, a) = \mathbf{1}\{u(t, a) \leq \alpha\}$.

A similar case can be made about models involving nuisance parameters. Suppose that $\theta = (\psi, \chi)$, where ψ is the potential parameter of interest and χ is a nuisance parameter. In these models it is sometimes possible to identify a minimal sufficient statistic (T, A) such that the distribution of $T|A$ is independent of χ and the distribution of A is independent of ψ , or depends on it in such a way that through observation of A alone no information can be extracted about ψ (Barndorff-Nielsen and Cox, 1994, Chapter 2). In such cases, a natural extension of the Conditionality Principle asserts that information about ψ should be interpreted via the conditional model $T|A$. As before, if the selection process interprets the sample information about ψ via the conditional distribution $T|A$, it is more natural

to define the selective density as

$$f_S(t, a; \theta) = \frac{f_A(a; \theta) f_{T|A}(t|a; \psi) p(t, a)}{\varphi(\psi; a)}$$

rather than

$$f_S(t, a; \theta) = \frac{f_A(a; \theta) f_{T|A}(t|a; \psi) p(t, a)}{\varphi(\psi, \chi)},$$

where we are using the same notational conventions as before.

An important setting where this argument applies is when inference is sought for a component of the canonical parameter in a full exponential family. Let the model be given by

$$f(y; \theta) = h(y) \exp \{ \psi s_1(y) + \chi^T s_2(y) - K(\theta) \}.$$

Standard arguments show that the distribution of $s_1(Y)|s_2(Y)$ is independent of χ , and even though the distribution of $s_2(Y)$ usually depends on ψ , power considerations based on Neyman-Pearson theory indicate that inference for ψ should be provided via the conditional distribution $s_1(Y)|s_2(Y)$ (Young and Smith, 2005, Chapter 7). Moreover, $s_2(Y)$ is uninformative for ψ in a sense made precise by Jørgensen (1994), so the Conditionality Principle applies. A second important class of examples with this structure that arise frequently in selective inference is the following.

Example 2.1.3. *Inference for winners.* Suppose that $Y_1, \dots, Y_m$ are independent observations with densities $f(y_i; \theta_i)$, $i = 1, \dots, m$, and suppose that we want to provide inference for the parameter which produces the largest observation, $\theta_I = \operatorname{argmax}\{y_i: i = 1, \dots, m\}$. Assume without loss of generality that $I = 1$, so that $T = Y_1$ and $A = (Y_2, \dots, Y_m)^T$. We can identify two natural sampling mechanisms that are consistent with selection of the first mean. In the first one, the whole vector $(Y_1, \dots, Y_m)^T$ is sampled until Y_1 is observed to be the maximum. The corresponding density of this generative process is

$$f_S(y_1, y_2, \dots, y_m; \theta) = \frac{\prod_{i=1}^m f(y_i; \theta_i) \mathbf{1}(I = 1)}{\mathbb{P}_\theta(Y_1 > Y_i \ \forall \ i > 1)}. \quad (2.1)$$

The second sampling mechanism is that in which $(Y_2, \dots, Y_m)^T$ is sampled from its unconditional distribution, and conditionally on its observed value, Y_1 is sampled until it exceeds $\max\{y_2, \dots, y_m\}$. The density of the data under this model is

$$f_S(y_1, y_2, \dots, y_m; \theta) = \frac{\prod_{i=1}^m f(y_i; \theta_i) \mathbf{1}(I = 1)}{\mathbb{P}_{\theta_1}(Y_1 > Y_i \ \forall \ i > 1 \mid y_2, \dots, y_m)}. \quad (2.2)$$

In the first model, conditioning on selection breaks the independence structure of the data. As a consequence, the observations from the non-selected parameters depend on θ_1 , and the marginal distribution of Y_1 depends on $\theta_2, \dots, \theta_m$. This makes manipulation of the likelihood function awkward and computationally expensive if m is large, despite the

apparent simplicity of the problem. One could nevertheless argue that it is appropriate to condition on the observed values of $Y_2, \dots, Y_m$, on the basis that they only depend mildly on the parameter of interest and the resulting inferences are simpler. However, if selection is used to determine whether θ_1 is the largest θ_i , and since the observed values of $Y_2, \dots, Y_m$ do not provide any direct information about this fact, preference of the sampling model (2.2) over (2.1) can be justified without resorting to computational considerations.

2.1.1 Ancillarity in selection models with nuisance parameters

The most intuitive definition of ancillarity in the presence of nuisance parameters requires the marginal distribution of the ancillary statistic A to be independent of ψ , which as we have just seen does not usually hold in selective distributions. Weaker definitions of ancillarity have been suggested that allow A to depend on ψ , but in such a way that if we only observed A we would not be able to extract any information about ψ . In this section we are going to show that two of these proposed notions of ancillarity are preserved after conditioning on selection, and can therefore be invoked to justify some of the claims made in the previous section in a more rigorous manner. Throughout this section we will assume that we can write $\theta = (\psi, \chi)$, that the parameter space can be written as $\Theta = \Theta_1 \times \Theta_2$, where Θ_1 and Θ_2 are the parameter spaces of ψ and χ , and that there exists a minimal sufficient statistic (T, A) for θ such that the distribution of $T|A$ is independent of χ . Instead of modifying the selective sampling distribution according to the nature of the sampling process, as in the previous section, we now assume that the selective distribution is $f_S(y; \theta) \propto f(y; \theta)p(y)$, and justify the change of inferential distribution differently.

In the previous setting, [Godambe \(1976\)](#) calls A ancillary if it is complete for χ for any fixed value of ψ . This is, if, for all ψ , whenever $\mathbb{E}_{\psi, \chi}[g(A)] = 0$ for all $\chi \in \Theta_2$ and some real-valued measurable function g , it must be the case that $\mathbb{P}_{\psi, \chi}\{g(A) = 0\} = 1$ for all $\chi \in \Theta_2$. We will refer to this type of ancillarity as G -ancillary. [Godambe \(1980\)](#) showed that if A is G -ancillary, then the distribution of T contains the same information about ψ as the distribution of $T|A$, so that the latter can be used for inference about ψ without any loss of information. It is easy to see that G -ancillarity is preserved after conditioning on selection events under minimal assumptions. We have the following result.

Proposition 1. If the selection mechanism is such that $\varphi(\psi; a) = \mathbb{P}_\psi(S \mid a) > 0$ for all possible (ψ, a) , then A is G -ancillary in the non-selective model if and only if it is G -ancillary in the selective model.

Proof. First of all, we have that, for any y_1, y_2 in the support of the selective distribution,

$$\frac{f_S(y_1; \theta)}{f_S(y_2; \theta)} = \frac{f(y_1; \theta) p(y_1)}{f(y_2; \theta) p(y_2)}$$

is independent of θ if and only if $f(y_1; \theta)/f(y_2; \theta)$ is independent of θ , which, by the minimal sufficiency of (T, A) in the non-selective regime, holds if and only if (T, A) is equal for y_1 and y_2 . This implies that (T, A) is also minimal sufficient for θ in the selective model. Secondly, we need to show that the distribution of $T|A$ in the selective model is independent of χ . This is trivial, as

$$f_S(t|a; \theta) = \frac{p(t, a)}{\varphi(\psi; a)} f(t|a; \psi).$$

Now, for A to be G -ancillary in the non-selective model it means that the condition $\mathbb{E}_{\psi, \chi}[g(A)] = 0$ for all χ implies that $\mathbb{P}_{\psi, \chi}\{g(A) = 0\} = 1$ for all χ , given an arbitrary ψ . However, note that

$$\mathbb{E}_{\psi, \chi}[g(A)|S] = \frac{\mathbb{E}_{\psi, \chi}[g(A)\varphi(\psi; A)]}{\varphi(\psi, \chi)},$$

so if $\mathbb{E}_{\psi, \chi}[g(A)|S] = 0$ we must have that $\mathbb{P}_{\psi, \chi}\{g(A)\varphi(\psi; A) = 0\} = 1$. With the assumption that $\varphi(\psi; a)$ is positive everywhere, this implies that $\mathbb{P}_{\psi, \chi}\{g(A) = 0\} = 1$, so A is G -ancillary in the selective model. The reciprocal claim follows by the same type of argument. $\square$

The second notion of ancillarity that we will consider stems from the field of noninformation theory, introduced by Barndorff-Nielsen in [Barndorff-Nielsen \(1976\)](#) and extended by [Jørgensen \(1994\)](#). Several types of ancillarity can be derived from this work. In the selective inference context, the most relevant one is M -ancillarity and extensions of thereof, all of which follow from the following general definition. If, for every possible pair (ψ, a) , the model $\{f_A(a; \psi, \chi) : \chi \in \Theta_2\}$ provides a *perfect fit* for the data a , then A is said to be ancillary ψ ([Jørgensen, 1994](#)). Different types of ancillarity can then be derived by adopting different definitions of perfect fit.

The rationale behind this definition follows from the following general principle: if two models provide a perfect fit for a given observation, then both are equally good at explaining the data, and we cannot assume that one is more likely to have generated the data than the other. Thus, in the previous setting, we cannot compare two values of ψ , ψ_1 and ψ_2 , say, only through observation of $A = a$, because the models $\{f_A(a; \psi_1, \chi) : \chi \in \Theta_2\}$ and $\{f_A(a; \psi_2, \chi) : \chi \in \Theta_2\}$ explain observation of a equally well.

In the case of M -ancillarity, perfect fit is defined as follows ([Barndorff-Nielsen, 1976](#)). A family of densities or mass functions $\{g(x; \lambda) : \lambda \in \Lambda\}$ is said to provide a perfect fit for

a given observation $x \in \mathcal{X}$ if there exists a $\lambda \in \Lambda$ such that x is an approximate mode of $g(x; \lambda)$. Formally, if

$$\text{for all } \varepsilon > 0 \text{ there exists a } \lambda \in \Lambda \text{ such that } g(x; \lambda) > (1 - \varepsilon) \sup_{\tilde{x} \in \mathcal{X}} g(\tilde{x}; \lambda). \quad (2.3)$$

One drawback of this definition is that it is not invariant under transformations of the data. For example, the family of exponential distributions does not provide a perfect fit for any positive observation, as all the densities have mode 0. However, one can easily show that after applying a log-transformation the family provides a perfect fit for any real observation. To remove this arbitrariness, we propose to change the definition and require instead that the previous condition holds under some change of variable. This produces a new type of ancillarity, which we term $\tilde{M}$ -ancillarity. We will assume that the distribution of A is continuous for simplicity. Extension to the discrete case is straightforward.

Definition 1. Let $\mathcal{A}$ be the sample space of A and assume it is open. A is said to be $\tilde{M}$ -ancillary for ψ if, for all $\psi \in \Theta_1$, there exists a smooth bijection $h : \mathcal{A} \rightarrow \mathcal{A}$ such that condition (2.3) holds for $h(A)$.

We then have the the following proposition.

Proposition 2. If the selection mechanism is such that $\varphi(\psi; a) = \mathbb{P}_\psi(S \mid a) > 0$ for all possible (ψ, a) , then A is $\tilde{M}$ -ancillary in the non-selective model if and only if it is $\tilde{M}$ -ancillary in the selective model.

Proof. The first part of the proof is the same as in Proposition 1. The marginal density of A given selection is

$$f_S(a; \theta) = \frac{\varphi(\psi; a)}{\varphi(\theta)} f_A(a; \theta).$$

Suppose that A is $\tilde{M}$ -ancillary for ψ in the non-selective model. Let ψ be arbitrarily fixed and define $Z_1 = h_1(A)$, and $Z_2 = h_2(Z_1)$, where h_1 is a bijection such that Z_1 satisfies the condition of Definition 1, and h_2 is a bijection with absolute Jacobian determinant $\varphi(\psi; h_1^{-1}(z_1)) |Jh_1^{-1}(z_1)|$, where $|Jh_1^{-1}(z_1)|$ denotes the absolute Jacobian determinant of $h_1^{-1}(z_1)$. By construction, the density of $Z_2|S$ is proportional to $f_{Z_1}(h_2^{-1}(z_2); \psi, \chi)$. By assumption, for any z_1 there exists a χ , $\hat{\chi}(z_1)$, such that z_1 is a mode of $f_{Z_1}(\cdot; \psi, \hat{\chi}(z_1))$. Hence, for all z_2 , $\hat{\chi}\{h_2^{-1}(z_2)\}$ is such that z_2 is a mode of $f_{Z_2}(\cdot|S; \psi, \hat{\chi}\{h_2^{-1}(z_2)\})$, which shows $\tilde{M}$ -ancillarity of A in the selective model. The converse can be proved in an analogous way. $\square$

The two forms of ancillarity considered in this section can be used to formally justify some of the arguments in the previous section. The simplest scenario is that described

in Example 2.1.3, where the non-selective distribution of the data is the product of m independent random variables with different underlying parameters. In that example, if the data from the non-selected parameters, $A = (Y_2, \dots, Y_m)^T$, is either complete or is such that the set of possible modes is the whole sample space, and if $\varphi(\psi; a)$ is positive everywhere, then A is respectively G or $\tilde{M}$ -ancillary for ψ in the selective model, and adherence to the Conditionality Principle would require providing inference for the selected parameter conditionally on the data from the non-selected parameters (i.e. according to model (2.2)). This occurs, for example, if $Y_i \sim N(\theta_i, 1)$, $\theta_i \in \mathbb{R}$. We might also consider a situation where $Y \sim N(\theta, I_n)$ follows a p -dimensional normal distribution for some $p > 1$, and inference is provided for a linear contrast $\psi = \eta^T \theta$. As we shall see, this problem arises in regression settings. In this case, if P is a matrix containing $p - 1$ linearly independent vectors orthogonal to η as columns, then $A = P^T Y$ can be easily seen to be both G and $\tilde{M}$ -ancillary for ψ without selection, and the same argument would apply.

Ancillarity in the presence of nuisance parameters is an intricate subject, and multiple definitions of it have been proposed, some of which are not preserved by conditioning on selection. A more complete examination of the relation between selection models and ancillarity would be interesting.

2.2 Bayesian inference

In this section we analyse the conceptual issues regarding the correct Bayesian treatment of selected parameters. The predominant viewpoint in the literature is that Bayesian inference should not be modified in the presence of selection. Quoting Dawid (1994):

Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data: that is, for a Bayesian, the face-value approach is fully valid, and no further adjustment for selection is required.

More informally, the fact that the statistician decides to focus on certain aspects of the data after it has been observed does not alter its sampling distribution, which should therefore remain unchanged.

As Dawid points out, the contrast between the Bayesian and the frequentist standpoints is somewhat paradoxical. In many situations, Bayesian analyses formally match, either exactly or approximately, face-value frequentist analyses, and it is universally agreed that

the latter methods are not valid in the presence of selection. Why, then, would such results be correct if reached by a Bayesian argument? Dawid argues that, certainly, whenever a frequentist approach is unreasonable, so is any Bayesian approach that provides similar answers. However, according to him, this does not reflect a fundamental issue about the Bayesian updating mechanism under selection. Rather, it is the result of a poor prior specification.

To understand why some prior distributions can be problematic under selection, consider the problem of providing inference for the mean of $Y \sim N(\theta, 1)$ only if $Y > 0$, and consider the standard class of conjugate priors for θ given by $\{N(0, \lambda^2) : \lambda > 0\}$. For a given λ , the posterior distribution of θ given $Y = y$ is normal with mean $\{\lambda^2/(1 + \lambda^2)\}y$ and variance $\lambda^2/(1 + \lambda^2)$. Small values of λ produce a shrinkage effect, pulling the posterior density of θ towards 0, while large values produce posterior inferences which are very similar to those provided by a face-value frequentist approach. Now, consider this problem from a frequentist perspective. If the true θ is significantly greater than zero, the selection bias will be very small and almost no adjustment is needed. Conversely, if θ is small, selection needs to be accounted for, as Y will tend to overestimate it. Therefore, the Bayesian analysis described before carried out with a large value of λ is appropriate only if we expect θ to be large, and will lead us astray if it is not. This is obvious, in a way: we would not adjust for selection if we believed a priori that θ is large and that an adjustment would not be necessary, and we would adjust for it if we suspected that θ is small and that therefore an adjustment would be required.

Of course, the fact that the prior choice is important and should not be taken lightly is not surprising. The key issue that the previous analysis highlights is that the impact of the prior on the analysis is stronger under selection, and, perhaps more importantly, that non-informative or weakly informative Bayesian analyses in these contexts are much more challenging than in non-selective settings. In the previous problem, if no selection takes place, lack of information about θ can be dealt with by taking a large, or even infinite, value of λ , as this minimises the influence of the prior on the analysis. In the presence of selection, however, this is no longer appropriate, as such theoretically “non-informative” priors do in fact entail critical information about the parameter, namely, its likelihood to lie in a region of the parameter space for which a selection adjustment would be appropriate.

However, even if we are comfortable with the implications of a certain prior in a selective problem, unadjusted Bayesian inference can still be problematic from a repeated-sampling viewpoint. Consider the model $Y \mid \theta \sim N(\theta, 0.2)$ and suppose that the prior distribution of θ is standard normal. In Figure 2.1 we have plotted the observed coverages of equal-tailed 0.9-credible intervals for fixed values of the true parameter in two different sampling

regimes: a selective one, where only the observations with $Y > 0$ were kept, and a non-selective one, where we constructed the intervals for all the sampled values of Y . The curves were obtained by simulation. We see that in the selective regime the intervals are very poorly calibrated for small values of θ , even for values with non-negligible selection probability (for example, when $\theta = -0.5$, the probability that $Y > 0$ is 0.13). By contrast, the coverage figures in the non-selective settings are much more stable. The instability observed in the selective regime is clearly undesirable, and, while a purist may argue that inferences are still well-calibrated once the coverage curve is averaged over the prior distribution, which in this case is $\pi(\theta | S) \propto \phi(\theta)\varphi(\theta)$, where ϕ is the standard normal density, most Bayesians would presumably agree that this behaviour is too extreme.

To avoid issues relating to selection, a natural option would be to construct the posterior distribution using only the information about θ provided by $Y | \{R = 1\}$, where as before R is the indicator variable of the selection event (Section 1.2). This way, the prior distribution is updated using only information which is not influenced by selection. Denoting the prior density by $\pi(\theta)$, the resulting posterior density would be

$$\pi(\theta | y) \propto \pi(\theta)f_S(y; \theta) \propto \frac{\pi(\theta)f(y; \theta)}{\varphi(\theta)}. \quad (2.4)$$

In the remainder of this work we will refer to posteriors of this form as selective posteriors. Inference based on selective posteriors allows the injection of prior information while avoiding potential problems arising from selection.

These type of posteriors have been considered by several authors, typically in settings where there is an explicit bias in the sampling mechanism of the data (Section 1.1). They are discussed by [Bayarri and DeGroot \(1987\)](#) and [Bayarri and Berger \(1998\)](#), and they also appear, in a different context, in [Bayarri and Berger \(2000\)](#), where they are referred to as “partial posterior densities” and are used for computing Bayesian p -values in a two-step procedure. As noted by [Harville \(2021\)](#), posterior (2.4) can also be thought of as following from a standard Bayesian updating of the modified prior $\pi^*(\theta) = \pi(\theta)/\varphi(\theta)$. This could be given the following interpretation: since unadjusted inference is very imprecise for values of θ with small selection probability, we assign a larger prior probability to those values to achieve a higher protection in those regions. In Figure 2.1 we have plotted the coverage curve produced by the selective posterior in the selective regime, which shows a much smoother behaviour than the unadjusted posterior.

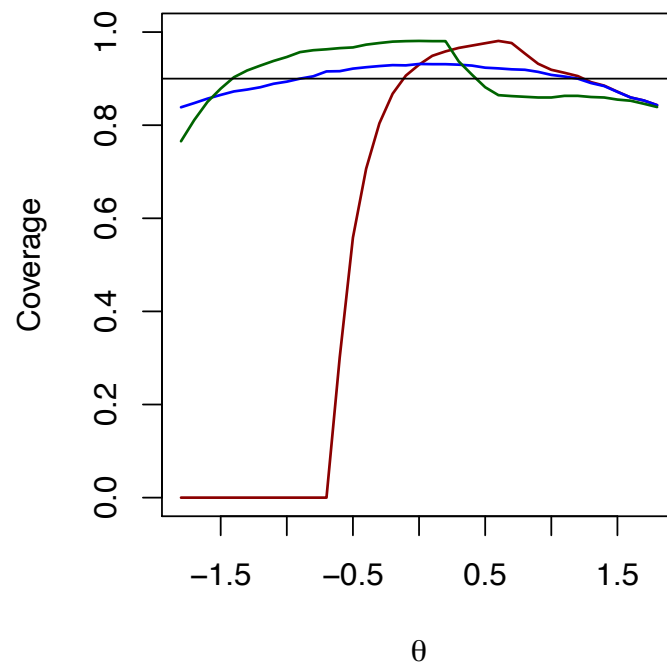


Figure 2.1: Coverage of credible intervals under repeated sampling from fixed values of θ derived from the unadjusted posterior (the red line corresponds to sampling with selection and the blue one without it), and from the selective posterior (green).

2.2.1 Fixed and random parameters

The belief that Bayesian inference is unaffected by selection has also been questioned by other authors on more conceptual grounds. Mandel and Rinott (2007, 2009) considered a sequence of binomial experiments in which only some of the inferences, corresponding to promising results in the setting considered, were reported. They made the observation that, if the sequence of parameters analysed had been sampled independently from a prior distribution, then no correction for selection was needed, in agreement with the classical stance. However, if the sequence of parameters had been generated by sampling one parameter from the prior and then successively testing it, then the appropriate inference needed a correction for selection. This idea was later extended and refined by Yekutieli (2012), who considered more general settings and linked these ideas to a Bayesian analogue of the false discovery rate.

According to Yekutieli, the correct Bayesian inference for a selected parameter depends on how the selection mechanism acts on the parameter space. Consider the joint sampling model of (θ, Y) and a selection function $p(y)$. Yekutieli calls θ *random* if the joint sampling scheme for the parameter and data is such that pairs (θ, Y) are sampled from their joint distribution until θ gets selected, and calls θ *fixed* if it is sampled from its marginal distribution, held fixed, and Y is sampled from the conditional distribution $Y \mid \theta$ until selection takes place.

As before, let R be the binary random variable that indicates if selection of the parameter under consideration has happened. If θ is random, its density given selection is $\pi(\theta \mid R = 1) \propto \pi(\theta)\varphi(\theta)$, whereas if it is fixed, its conditional density is unchanged, $\pi(\theta \mid R = 1) = \pi(\theta)$. In the case of the data, its conditional density given θ and the selection event is $f_S(y; \theta) = f(y; \theta)p(y)/\varphi(\theta)$ in both cases. Thus, the posterior distribution for a random parameter is

$$\pi(\theta \mid y) \propto \frac{\pi(\theta)\varphi(\theta)f(y; \theta)}{\varphi(\theta)} = \pi(\theta)f(y; \theta), \quad (2.5)$$

while for a fixed parameter it is given by

$$\pi(\theta \mid y) \propto \frac{\pi(\theta)f(y; \theta)}{\varphi(\theta)}. \quad (2.6)$$

Hence, Bayesian inference for a selected random parameter is unaffected by selection, but for a fixed parameter the posterior needs to be adjusted, and it is formally constructed by attaching the prior density to the selective likelihood. Yekutieli argues that Dawid's analysis follows from the implicit assumption that the parameter is random.

To elucidate the difference between these two regimes let us consider the following ex-

ample, which is a slight variation of the first example in Yekutieli (2012). Suppose that θ represents the average academic ability of the students in a given population. Each student takes an exam and gets a grade Y_i , and the students whose grade exceeds a given threshold t are admitted to university. If a person in the university, who has access to the grades obtained by the admitted students and to prior knowledge about θ in the form of historical data, say, wants to estimate the average academic ability in the population, they should carry out the analysis treating θ as a fixed parameter. If, instead, θ is a student-specific parameter that represents the student's academic ability, and the person in the university wants to estimate each of the academic abilities of the admitted students, then each estimated θ should be treated as a random parameter.

The previous analysis, however, does not cover situations with non-informative priors. It is argued by Yekutieli (Section 2.2) that, in cases where a non-informative prior would have been used without selection, a non-informative prior should also be used with selection, regardless of whether θ is random or fixed. The reason given is that, just as in the non-selective regime, all the information about the parameter in the selective problem comes from the data. Yekutieli further suggests to use the same non-informative prior in the selective and non-selective models, because the decision of providing inference for a particular parameter should not alter the prior.

We agree with the first assertion but not with the second one, that is, we believe that the non-informative prior used in the selective model should be different from the one that would be used in the non-selective one. Non-informative priors are typically designed to achieve certain properties with respect to the sampling model of the data. In a selection model, both in the random and fixed parameter regimes, the sampling distribution of the data is the same (it has density or mass function $f_S(y; \theta)$), so the non-informative prior should also be the same in both cases. Furthermore, since the sampling distribution of the data depends on the selection mechanism, we argue that any appropriate choice of non-informative prior should also depend on selection. Consider, for example, a location model $f(y; \theta) = g(y - \theta)$. The standard non-informative prior for θ in absence of selection is $\pi(\theta) \propto 1$, which enjoys several interesting properties: it is Jeffreys, minimises the prior influence, and produces inferences with a valid frequentist interpretation. However, all these nice properties no longer hold if inference for θ is only provided for certain values of y , as the selective model $\{g(y - \theta)p(y)/\varphi(\theta) : \theta \in \Theta\}$ is not a location model unless $p(y) \propto 1$. We expand on these ideas in Chapter 4.

As for informative priors, while the posterior densities (2.5) and (2.6) are formally correct given the respective sampling mechanisms, it is not clear that a parameter can be labelled as random or fixed without explicit consideration of this sampling process, which may not be well-defined in some cases. We therefore suggest to use the adjusted posterior to avoid

any potential bias that may arise due to selection. We note, however, that the distinction between random and fixed parameters is clarifying, in that it captures the two types of bias that can arise because of selection. The first occurs in the sample space, when the same parameter is analysed multiple times: if some of the analyses are not reported because the parameter did not appear to be significant in light of the data, the published reports about it will on average overstate this significance. The second one is a bias in the parameter space: parameters that are in some way interesting are more likely to be analysed than those which are not.

Chapter 3

Some results about selection models

In this chapter we provide a theoretical analysis about some aspects of selection models. First, we consider two point estimators for the mean of a selected normal population and study their asymptotic properties. Then, we show that a class of regular one-dimensional selective models can be asymptotically approximated by selective normal-location models in a neighbourhood of the true parameter.

3.1 Point estimation

Consider a selection model involving a p -dimensional normal observation $Y \sim N(\theta, \Sigma)$ and a selection function $p(y)$, where $\theta \in \mathbb{R}^p$ is unknown and $\Sigma \in \mathbb{R}^{p \times p}$ is a known positive definite matrix. The selective density of this model is given by

$$f_S(y; \theta) = \frac{\exp \left\{ -(1/2)(y - \theta)^T \Sigma^{-1} (y - \theta) \right\} p(y)}{(2\pi)^{p/2} |\Sigma|^{1/2} \varphi(\theta)}, \quad \varphi(\theta) = \mathbb{E}_\theta[p(Y)],$$

and may be written in exponential family form as

$$f_S(y; \theta) = h(y) \exp \left\{ \theta^T s(y) - A(\theta) \right\},$$

where $h(y) = p(y) \exp \{ -(1/2)y^T \Sigma^{-1} y \}$, $s(y) = \Sigma^{-1} y$, and $A(\theta) = (1/2)\theta^T \Sigma^{-1} \theta + \log \varphi(\theta)$. In particular, the cumulant generating function of Y under θ is given by

$$K(t) = A(\theta + \Sigma t) - A(\theta) = \theta^T t + (1/2)t^T \Sigma t + \log \frac{\varphi(\theta + \Sigma t)}{\varphi(\theta)}.$$

We are going to consider two natural estimators of θ in this model. First, the maximum likelihood estimator (MLE), defined as the maximiser the selective likelihood. It satisfies the equation

$$y = \hat{\theta} + \Sigma \nabla \log \varphi(\hat{\theta}),$$

where, for a smooth function $g: \mathbb{R}^p \rightarrow \mathbb{R}$, $\nabla g: \mathbb{R}^p \rightarrow \mathbb{R}^p$ denotes its gradient. This equation needs to be solved numerically except in trivial cases.

The other estimator that we are going to consider is the maximum ordinate estimator (MOE). Maximum ordinate estimators were introduced by Höglund (1974) and are defined, for a given y , as the value of the parameter under which the distribution has mode y . For selective normal models with a selection function which is smooth in $\mathbb{R}^p$, this estimator is unbiased for θ . Furthermore, it can be written out explicitly, which makes it easier to implement than the MLE when p is large.

If the selection function is continuously differentiable around y , we can easily see that the MOE is given by

$$\tilde{\theta} = y - \Sigma \nabla \log p(y).$$

Note that, if $p(y)$ is constant around y (for example, if $p(y) = \mathbf{1}(y \in E)$ for some event E), then $\tilde{\theta} = y$, which is the standard estimator of θ without selection. The proof that $\tilde{\theta}$ is unbiased when $p(y)$ is smooth is straightforward. Simply note that we can write it as $\tilde{\theta} = \theta - \Sigma \nabla \log f_S(y; \theta)$, and that the second term has null expectation under θ , as

$$\int_{-\infty}^{\infty} \left\{ \frac{\partial}{\partial \tilde{y}_i} \log f_S(\tilde{y}; \theta) \right\} f_S(\tilde{y}; \theta) d\tilde{y}_i = \int_{-\infty}^{\infty} \frac{\partial}{\partial \tilde{y}_i} f_S(\tilde{y}; \theta) d\tilde{y}_i = 0$$

for all $i = 1, \dots, p$. Furthermore, since $\Sigma^{-1}Y$ is the natural sufficient statistic of a full exponential family, Y is sufficient and complete, so $\tilde{\theta}$ is optimal among the unbiased estimators of θ by the Lehmann–Scheffé Theorem.

In some situations we may have access to two independent observations, $Y_1 \sim N(\theta, n_1^{-1}\Sigma)$ and $Y_2 \sim N(\theta, n_2^{-1}\Sigma)$, and a selection event of the form $Y_1 \in E$ for some pre-specified $E \subseteq \mathbb{R}^p$. This model arises when a random sample from $N(\theta, \Sigma)$ of size n_1 is available for selection, and a second sample of size n_2 is collected after selection, should it take place. Letting $n = n_1 + n_2$, $\gamma = n_1/n$, and $Y = \gamma Y_1 + (1 - \gamma)Y_2$, this model can be reduced by sufficiency to one involving a single observation $Y \sim N(\theta, n^{-1}\Sigma)$ and selection function $p(y) = \mathbb{P}(y + W \in E)$, where $W \sim N(0_p, n^{-1}\{\gamma^{-1} - 1\}\Sigma)$ is independent of Y and $0_p \in \mathbb{R}^p$ is a vector of zeroes. To see this, simply define $W = (1 - \gamma)(Y_1 - Y_2)$, which follows the claimed distribution and is independent of Y , as they are uncorrelated and normal, and note that $Y_1 = Y + W$. This analysis shows that, in a normal location model, selection based on a γ -split of the data is equivalent to an additive randomised selection scheme with

randomisation noise W , as both selection strategies produce the same selection function.

In these models, the MOE estimator follows from Rao-Blackwellization of the trivially unbiased estimator $Y_2 = Y - \{\gamma/(1-\gamma)\}W$, and can be easily approximated via simulation. Since $p(y) = \mathbb{P}\{N(y, n^{-1}\{\gamma^{-1} - 1\}\Sigma) \in E\}$, we have that

$$\nabla \log p(y) = \frac{n\gamma}{1-\gamma} \Sigma^{-1} \mathbb{E}[T - y \mid T \in E], \quad T \sim N(y, n^{-1}\{\gamma^{-1} - 1\}\Sigma),$$

so

$$\tilde{\theta} = \frac{1}{1-\gamma} \{y - \gamma \mathbb{E}[T]\}.$$

Therefore, for any given y , $\tilde{\theta}$ can be approximated by

$$\tilde{\theta}^* = \frac{1}{1-\gamma} \left\{ y - \frac{\gamma}{M} \sum_{i=1}^M T_i \right\},$$

where $T_1, \dots, T_M$ are independent samples from the distribution $N(y, n^{-1}\{\gamma^{-1} - 1\}\Sigma)$ truncated to E , which may be generated via rejection sampling. Note also that, since the Monte Carlo approximation of $\mathbb{E}[T]$ is unbiased, $\tilde{\theta}^*$ is unbiased for θ for any given simulation size M .

In the one-dimensional case, $p = 1$, the two most common forms of selection are one-sided selection, $Y_1 > t$, and two-sided selection, $|Y_1| > t$, for some threshold $t \in \mathbb{R}$. Assuming that $\Sigma = 1$ for simplicity, the data-splitting models described previously have selection function $p(y) = \Phi[\{n\gamma/(1-\gamma)\}^{1/2}(y-t)]$ for one-sided selection, and $p(y) = 1 + \Phi[\{n\gamma/(1-\gamma)\}^{1/2}(y-t)] - \Phi[\{n\gamma/(1-\gamma)\}^{1/2}(y+t)]$ for two-sided selection. In addition, the respective selection probabilities are $\varphi(\theta) = \Phi\{(\gamma n)^{1/2}(\theta-t)\}$ and $\varphi(\theta) = 1 + \Phi\{(\gamma n)^{1/2}(\theta-t)\} - \Phi\{(\gamma n)^{1/2}(\theta+t)\}$. Therefore, in the first case the MLE equation is

$$y = \hat{\theta} + \left(\frac{\gamma}{n}\right)^{1/2} h_1 \left\{ (\gamma n)^{1/2}(\hat{\theta} - t) \right\},$$

where $h_1(x) = \phi(x)/\Phi(x)$, and the MOE is given, for $\gamma < 1$, by

$$\tilde{\theta} = y - \frac{1}{n^{1/2}} \left(\frac{\gamma}{1-\gamma} \right)^{1/2} h_1 \left\{ \left(\frac{n\gamma}{1-\gamma} \right)^{1/2} (y-t) \right\}.$$

For two-sided selection we have

$$y = \hat{\theta} + \left(\frac{\gamma}{n}\right)^{1/2} \frac{\phi\{(\gamma n)^{1/2}(\hat{\theta} - t)\} - \phi\{(\gamma n)^{1/2}(\hat{\theta} + t)\}}{1 + \Phi\{(\gamma n)^{1/2}(\hat{\theta} - t)\} - \Phi\{(\gamma n)^{1/2}(\hat{\theta} + t)\}}$$

and

$$\tilde{\theta} = y - \frac{1}{n^{1/2}} \left(\frac{\gamma}{1-\gamma} \right)^{1/2} \frac{\phi[\{n\gamma/(1-\gamma)\}^{1/2}(y-t)] - \phi[\{n\gamma/(1-\gamma)\}^{1/2}(y+t)]}{1 + \Phi[\{n\gamma/(1-\gamma)\}^{1/2}(y-t)] - \Phi[\{n\gamma/(1-\gamma)\}^{1/2}(y+t)]}.$$

These estimators have been thoroughly studied in the biostatistics literature for the problem of estimating the mean (typically a treatment effect) that provides the largest observation out of m normal samples with common variance (see e.g. [Cohen and Sackrowitz, 1989](#); [Shen, 2001](#); [Stallard and Todd, 2005](#); [Stallard et al., 2008](#)). Formally, this model involves m independent observations $Y_i \sim N(\theta_i, n^{-1})$, $i = 1, \dots, m$, and inference is provided for θ_I , where $Y_I = \max\{Y_1, \dots, Y_m\}$. In addition, sometimes an extra set of samples from the selected effect is collected after selection. Recall from the previous chapter that in this model there are strong arguments for conditioning on the data of the non-selected populations. If we carry out such conditioning, the resulting model is a one-dimensional model with selection event $Y_I > t$, with $t = \max\{y_i: i \neq I\}$. Furthermore, providing inference for the means of the top k observations rather than just for that of the largest one does not entail any extra complications. After conditioning on the data from the non-selected means, the observations from the selected ones are independent, so the full model reduces to k independent one-dimensional models, which can be handled separately. In some situations a control sample Z may be available and the selection rule compares the differences $\{Y_i - Z: i = 1, \dots, m\}$. Since Z is also uninformative about the selected mean, the arguments of the previous chapter apply and inference would condition on its observed value, so that the selection event becomes again a one-sided interval.

In Figure 3.1 we have plotted $\hat{\theta}$ and $\tilde{\theta}$ as a function of the data y for different values of γ . For one-sided selection the estimators are nearly identical, but for two-sided selection they exhibit a very different behaviour. The MLE is an increasing function of y , as one would expect, since the selective model is stochastically increasing in θ , while the unbiased estimator has a very sharp behaviour near $y = 0$, where the selection probability is the lowest. Furthermore, it is not an increasing function of y . To see why this could happen, suppose that $n_1 = n_2 = 1$, $t = 2$, and $y = 0.5$. Given this value of y and the fact that $|Y_1| > t$, it is far more likely that $Y_1 > t$ than $Y_1 < -t$. Specifically,

$$\frac{\mathbb{P}(Y_1 > t \mid |Y_1| > t, y)}{\mathbb{P}(Y_1 < -t \mid |Y_1| > t, y)} = \frac{\mathbb{P}(Y_1 > t \mid y)}{\mathbb{P}(Y_1 < -t \mid y)} = \frac{\Phi\{2^{1/2}(y-t)\}}{\Phi\{-2^{1/2}(y+t)\}} \approx 83,$$

as $Y_1 \mid y \sim N(y, 1/2)$. Since $y_2 = 2y - y_1$, if $y_1 > t$ we have that $y_2 < 2y - t = -1$, so the values of y_1 and y_2 differ considerably. Since y_2 has been sampled without selection, and therefore contains more information about θ , it would appear more adequate to assign a higher inferential weight to the second observation, pulling the estimate towards y_2 . Thus, the behaviour observed in the second plot is not entirely unreasonable.

The case $\gamma = 1$ is tricky from an inferential viewpoint, because as $n \rightarrow \infty$ the model becomes highly non-identifiable on the lower end of the parameter space (all small enough values θ produce data points which are very close to the selection boundary). [Stallard et al. \(2008\)](#) show that in this case $\mathbb{E}_\theta[|\hat{\theta}| \mid S] = \infty$ for any true θ , so the MLE is highly unstable even when the true mean is large. Furthermore, they also show that unbiased estimation is not possible in this regime. Clearly, the MOE estimator does not provide a more reliable option in this scenario, as in this case it is simply equal to the unadjusted estimator Y , which can be considerably biased.

To conclude this section, in the following lemma we establish some basic results concerning the asymptotic behaviour of these estimators under one-sided selection. In what follows, all $O_p(\cdot)$ and $o_p(\cdot)$ statements are with respect to the selective distribution of the data. The proof of the result can be found in the appendix.

Lemma 1. For each $n \in \mathbb{N}$ consider the selective model with data $Y_n \sim N(\theta, n^{-1})$ and selection function $p_n(y) = \Phi[\{n\gamma/(1-\gamma)\}^{1/2}(y - t_n)]$ for $\gamma > 1$ and $p_n(y) = \mathbf{1}(y > t_n)$ for $\gamma = 1$, where $(t_n)_{n \in \mathbb{N}} \in \mathbb{R}$ is a sequence of thresholds. Denote the true generating parameter in the n -th model by θ_n , the selection event by S_n , and denote the MLE and MOE estimators by $\hat{\theta}_n$ and $\tilde{\theta}_n$. Then:

1. If $\gamma = 1$ for all n and $\theta_n < t_n - \varepsilon$ for all sufficiently large n and some $\varepsilon > 0$, both $\tilde{\theta}_n$ and $\hat{\theta}_n$ are inconsistent. Furthermore, $\tilde{\theta}_n - t_n$ converges in probability to 0 and

$$\frac{\mathbb{P}_{\theta_n}(\hat{\theta}_n \leq t_n - x \mid S_n)}{\mathbb{P}\{\text{Exp}(1) \leq x^{-1}(t_n - \theta_n)^{-1}\}} \xrightarrow{n \rightarrow \infty} 1 \quad \text{for all } x > 0,$$

where $\text{Exp}(1)$ denotes the exponential distribution with rate 1. In particular, $\mathbb{P}_{\theta_n}(\hat{\theta}_n \leq t_n - x \mid S_n)$ is bounded away from zero for every $x > 0$.

2. If $\gamma = 1$ and $\varphi_n(\theta_n)$ is bounded away from zero, $|\hat{\theta}_n - \theta_n| = O_p(n^{-1/2})$ and $|\tilde{\theta}_n - \theta_n| = O_p(n^{-1/2})$.
3. If $\gamma < 1$, $|\hat{\theta}_n - \theta_n| = O_p(n^{-1/2})$ and $|\tilde{\theta}_n - \theta_n| = O_p(n^{-1/2})$ for any sequence $(\theta_n)_{n \in \mathbb{N}} \in \mathbb{R}$.

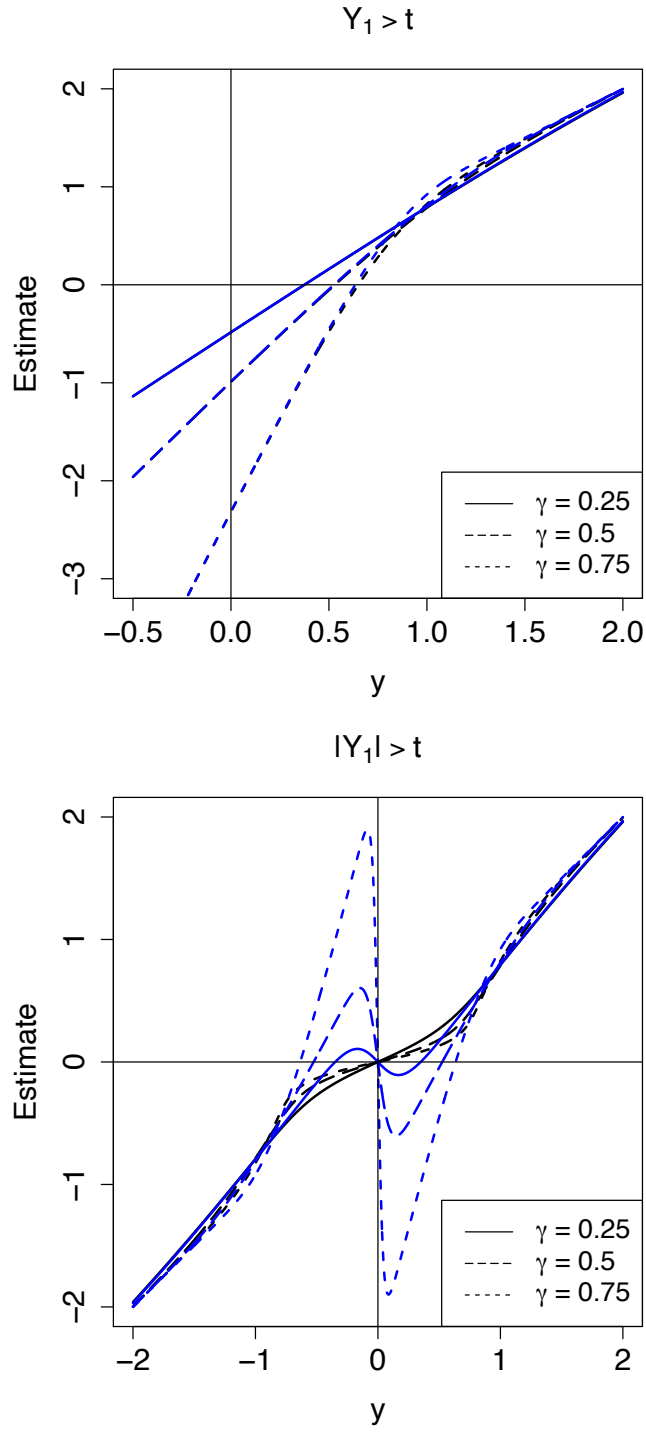


Figure 3.1: MLE (black) and MOE (blue) as a function of y for one and two-sided selection, with $n = 10$, $t = 2n_1^{-1/2}$, and $\gamma = 0.25, 0.5, 0.75$.

3.2 Asymptotic selective normality

Much of the existing theory on selective inference is restricted to normal models. It is therefore convenient to study whether the guarantees of the inferential procedures designed for these settings are approximately valid in a broader class of problems. In the absence of selection, if the data consists of a random sample from a regular model and the sample size is large, the resulting inferential problem is approximately equivalent to a normal inference problem with unknown mean and known variance in a $n^{1/2}$ -neighbourhood of the true parameter. This was elegantly formalised by [Le Cam \(1960\)](#) via the concept of Local Asymptotic Normality (LAN); see also Chapter 7 of [van der Vaart \(1998\)](#). Local Asymptotic Normality lies at the heart of many well-known asymptotic results in statistics, including the asymptotic normality of the MLE and the Bernstein-von Mises Theorem, which establishes the asymptotic behaviour of the posterior distribution under a sensible prior. Unfortunately, when the data is subject to some form of selection, asymptotic normality need not hold, and inferences based on a limiting normal approximation can be misleading. In this section we explore an extension of the LAN framework to selection problems.

Let us consider a sequence of selection models defined over a random sample from a one-dimensional family. For each $n = 1, 2, \dots$, we assume the model involves n independent observations $Y_1, \dots, Y_n$ from some parametric family $F(y; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}$, and that this sample is subject to a selection mechanism encoded as a selection function $p_n(y_1, \dots, y_n)$, which typically varies with the sample size. The corresponding selection probability will be denoted by $\varphi_n(\theta)$, and the selection event by S_n . In some cases the selection function will depend only on a subset of $n_1 < n$ observations, without loss of generality assumed to be the first n_1 , so that we may write the selection function as $p_n(y_1, \dots, y_n) \equiv p_n(y_1, \dots, y_{n_1})$. We will write $\gamma = n_1/n$ for the fraction of observations used for selection, and will assume that γ stays approximately fixed as n increases. Cases with $\gamma < 1$ correspond to situations where either the original sample has been split or we have been able to obtain an extra set of observations after selection. We will denote the true parameter by θ_0 . Depending on the limiting behaviour of the selection probability at θ_0 we can have one of the following three possibilities:

- **Large-deviations** setting: $\varphi_n(\theta_0) \rightarrow 0$ as $n \rightarrow \infty$. When $\gamma = 1$ this is a complex regime to analyse, because the behaviour of the model for large n is dominated by the tail behaviour of the distribution, about which we typically have little information. When further data is available for inference, $\gamma < 1$, the information provided by the observations employed on the selection stage is typically negligible relative to that provided by the unused ones, and asymptotically we end up having a standard IID

situation with $(1 - \gamma)n$ observations.

- **Moderate-deviations** setting: $a < \varphi_n(\theta_0) < b$ for some $0 < a < b < 1$ for all large enough n . In this regime, the effect of selection on the model is non-negligible and needs to be corrected for even when n is large. However, it admits a simpler asymptotic analysis than the large-deviations setting.
- **Small-deviations** setting: $\varphi_n(\theta_0) \rightarrow 1$ as $n \rightarrow \infty$. It occurs when the true parameter is well within the selection region. This is the easiest setting to analyse, because as n increases the effect of selection on the model vanishes and can typically be ignored for practical purposes.

Before deriving the asymptotic behaviour of selective models, we review a classical theorem for non-selective models, of which our result can be regarded a generalisation. We state the theorem assuming smoothness of the likelihood, but this assumption can be weakened.

Theorem 1. ([van der Vaart, 1998](#), Theorem 7.2) Let $Y_1, \dots, Y_n$ be a random sample from a distribution with density or mass function $f(y_i; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}$, with Θ open. Suppose that the map $\theta \rightarrow f(y_i; \theta)^{1/2}$ is continuously differentiable for every observable y_i , and that the Fisher information $i(\theta) = \mathbb{E}_\theta[-\{\partial^2 / \partial^2 \theta\} \log f(Y_i; \theta)]$ exists and is continuous in θ . Then, for every converging sequence $u_n \rightarrow u$,

$$\log \frac{f(Y_1, \dots, Y_n; \theta_0 + n^{-1/2} u_n)}{f(Y_1, \dots, Y_n; \theta_0)} = ui(\theta_0)Z_n - \frac{u^2}{2}i(\theta_0) + o_p(1), \quad (3.1)$$

where $Z_n \xrightarrow{d} N(0, i(\theta)^{-1})$.

When expansion (3.1) holds we say that the sequence of models is Locally Asymptotically Normal. The rationale behind the asymptotic expansion of the log-likelihood ratio is the following. Consider the model $\{N(u, i(\theta_0)^{-1}) : u \in \mathbb{R}\}$, and denote the density of a $N(u, i(\theta_0)^{-1})$ random variable by $g(z; u)$. Then, for $Z \sim N(0, i(\theta_0)^{-1})$,

$$\log \frac{g(Z; u)}{g(Z; 0)} = ui(\theta_0)Z - \frac{u^2}{2}i(\theta_0),$$

which is asymptotically equivalent, in distribution, to the dominant term of (3.1). In other words, in a $n^{1/2}$ -neighbourhood of the true parameter, the likelihood ratio process of the original model is asymptotically equivalent to the likelihood ratio process of a normal-location model with known variance $i(\theta_0)^{-1}$. We will not detail the implications of Local Asymptotic Normality in this work, but the key idea is that, whenever the LAN expansion holds, the asymptotic properties of inferential procedures for the original model

can be explained by analysing their behaviour when applied to the limiting normal-location problem. For example, the largest power achievable by a testing procedure in the original model is the largest one achievable in the normal model; we refer the reader to Chapters 7 and 9 of [van der Vaart \(1998\)](#) for explicit results.

In view of this, it is natural to ask under which conditions a similar expansion holds for selective models. Consider now the selective model with two independent observations $Z_1 \sim N(\gamma^{1/2}u, i(\theta_0)^{-1})$ and $Z_2 \sim N(\{1 - \gamma\}^{1/2}u, i(\theta_0)^{-1})$, and selection function $p(z_1)$ depending only on the first observation, where $\gamma \in [0, 1]$ is known and $u \in \mathbb{R}$ is unknown. Denoting by $g_S(z_1, z_2; u) \propto \phi\{i(\theta_0)^{1/2}(z_1 - \gamma^{1/2}u)\}\phi\{i(\theta_0)^{1/2}(z_2 - (1 - \gamma)^{1/2}u)\}p(z_1)$ the selective density, we have that the selective log-likelihood ratio can be written as

$$\log \frac{g_S(Z_1, Z_2; u)}{g_S(Z_1, Z_2; 0)} = ui(\theta_0)\{\gamma^{1/2}Z_1 + (1 - \gamma)^{1/2}Z_2\} - \frac{u^2}{2}i(\theta_0) + \log \frac{\varphi(0)}{\varphi(u)},$$

where $\varphi(u) = \mathbb{E}_u[p(Z_1)]$. Note that, while the log-likelihood ratio of a normal model follows itself a normal distribution, the log-likelihood ratio of a selective normal model follows a selective normal distribution.

In the following result we provide conditions under which the log-likelihood ratio process of a generic selective model admits an asymptotic expansion of the previous form. The proof can be found in the appendix. Regularity conditions aside, the most relevant assumption of the theorem is that the selection probability at the true parameter does not vanish. This allows us to work in moderate deviation regions, where standard asymptotic theory can be applied.

Theorem 2. Let $\gamma \in [0, 1]$ be fixed and consider a sequence of selective models of the form described before in which the selection function depends only on the first $n_1 = \lfloor \gamma n \rfloor$ observations, where $\lfloor \cdot \rfloor$ denotes the integer part, and let $n_2 = n - n_1$. Assume that the parameter space Θ is open and connected, and that the non-selective log-likelihood function of a single observation, $l_1(\theta) = \log f(y_1; \theta)$, is three times continuously differentiable for every observable y_1 . Furthermore, for a fixed true $\theta_0 \in \Theta$, assume that $\varphi_n(\theta_0)$ is bounded away from zero, the Fisher information $i(\theta_0) = -\mathbb{E}_{\theta_0}[l_1''(\theta_0)]$ exists, define the selection function

$$\tilde{p}_{n_1}(z_1; \theta_0) = \mathbb{E}_{\theta_0}[p_n(Y_1, \dots, Y_{n_1}) \mid l'_{n_1}(\theta_0) = n_1^{1/2}i(\theta_0)z_1],$$

where $l_{n_1}(\theta) = \sum_{i=1}^{n_1} \log f(y_i; \theta)$, and assume that the following conditions hold:

1. For all n_1 , $\mathcal{P}_{n_1} = \{z_1 \in \mathbb{R} : \tilde{p}_{n_1}(z_1; \theta_0) > 0\}$ is the union of at most k intervals for some fixed $k \in \mathbb{N}$;
2. For all n_1 , $\tilde{p}_{n_1}(z_1; \theta_0)$ is continuously differentiable and has a bounded derivative in the interior of $\mathcal{P}_{n_1}$, and satisfies $\int_{\mathcal{P}_{n_1}} |\tilde{p}'_{n_1}(z_1; \theta_0)| dz_1 = O(1)$;

3. $\mathbb{E}_{\theta_0}[\exp\{xl'_1(\theta_0)\}] < \infty$ for all $x \in \mathbb{R}$;
4. $\mathbb{E}_{\theta_0}[\exp\{xl''_1(\theta_0 + y)\}] < \infty$ for all x, y in a non-empty ball centred at 0;
5. $\mathbb{E}_{\theta_0}[l'''_1(\theta_0 + x)] < \infty$ for all x in a non-empty ball centred at 0.

Then, for any convergent sequence $u_n \rightarrow u$, the selective log-likelihood ratio admits the expansion

$$\log \frac{f_S(Y_1, \dots, Y_n; \theta_0 + n^{-1/2}u_n)}{f_S(Y_1, \dots, Y_n; \theta_0)} = ui(\theta_0)\{\gamma^{1/2}Z_{n_1} + (1-\gamma)^{1/2}Z_{n_2}\} - \frac{u^2}{2}i(\theta_0) + \log \frac{\tilde{\varphi}_{n_1}(0)}{\tilde{\varphi}_{n_1}(u)} + o_p(1)$$

conditionally on selection, where:

1. $G_{n_1}(Z_{n_1}) \xrightarrow{d} U(0, 1)$, with $G_{n_1}(x)$ the distribution function of the selective distribution of $Z_1 \sim N(0, i(\theta_0)^{-1})$ with selection function $\tilde{p}_{n_1}(z_1; \theta_0)$;
2. $Z_{n_2} \xrightarrow{d} N(0, i(\theta_0)^{-1})$;
3. $\tilde{\varphi}_{n_1}(u) = \mathbb{E}[\tilde{p}_{n_1}(Z_1; \theta_0)]$, with $Z_1 \sim N(u, i(\theta_0)^{-1})$.

Note that, since the selection rule usually depends on the sample size, the dominant term of the likelihood ratio does not, in general, converge in distribution to any random variable. Rather, it approximates the likelihood ratio of a sequence of normal selection models. As a consequence, standard limit of experiments theory does not apply to this problem, and a formal justification is needed to apply those results to this setting. Such extensions are not derived in this work, but constitute an important direction for future research. As an empirical check, below we compare the asymptotic distribution of a selective MLE to the asymptotic distribution predicted by the theorem.

The simplest class of models that satisfy the regularity conditions of Theorem 2 are exponential family models. Let the density or mass function of the observations be given by

$$f(y_i; \theta) = h(y_i) \exp\{\theta s(y_i) - A(\theta)\}.$$

Then $l'_1(\theta) = s(Y_1) - A'(\theta)$, $l''_1(\theta) = -A''(\theta)$, and $l'''_1(\theta) = A'''(\theta)$, and by standard properties of exponential families we have that $\mathbb{E}_\theta[xl'_1(\theta)] = \exp\{A(\theta + x) - A(\theta) - xA'(\theta)\}$, which verifies assumption 3. Assumptions 4 and 5 are trivially satisfied. Now, suppose that the selection event is of the form $\sum_{i=1}^{n_1} s(Y_i) > t_{n_1}$ for some sequence of thresholds t_{n_1} . This type of selection corresponds to carrying out a uniformly most powerful test of a one-sided hypothesis about θ , using the first n_1 observations (similarly, we might also consider two-sided tests). In this case we would have $\tilde{p}_{n_1}(z_1; \theta_0) = \mathbf{1}\{z_1 > i(\theta_0)^{-1}n_1^{-1/2}(t_{n_1} - n_1A'(\theta_0))\}$, where $i(\theta_0) = A''(\theta_0)$, which clearly satisfies conditions 1 and 2.

Consider now the problem of estimating θ in a model satisfying the conditions of Theorem 2. The MLE of the local parameter $u = n^{-1/2}(\theta - \theta_0)$ in the selective model can be expressed as

$$\hat{u} = \operatorname{argmax}_u \log \frac{f_S(Y_1, \dots, Y_n; \theta_0 + n^{-1/2}u)}{f_S(Y_1, \dots, Y_n; \theta_0)}.$$

We might therefore expect the distribution of $\hat{u}$ to be close to that of

$$u^* = \operatorname{argmax}_u ui(\theta_0) \{ \gamma^{1/2} Z_{n_1} + (1 - \gamma)^{1/2} Z_{n_2} \} - \frac{u^2}{2} i(\theta_0) + \log \frac{\tilde{\varphi}_{n_1}(0)}{\tilde{\varphi}_{n_1}(u)}$$

when n is large, with Z_{ni} having the asymptotic distributions given in the theorem. In Figure 3.2 we compare the cumulative distribution functions of $\hat{u}$ and u^* when the data is exponentially distributed. The observations are $Y_i \sim \text{Exp}(\theta)$, where θ is the rate parameter and, for each of the indicated values of γ and n , we considered the selection event $n_1^{-1} \sum_{i=1}^{n_1} Y_i > 1$, and we chose the true parameter so that $\varphi_n(\theta_0) = 0.5$. All the distribution functions were computed by simulation with a large sample size. As we can see, the actual distribution of the MLE is very close to the predicted one in all the cases.

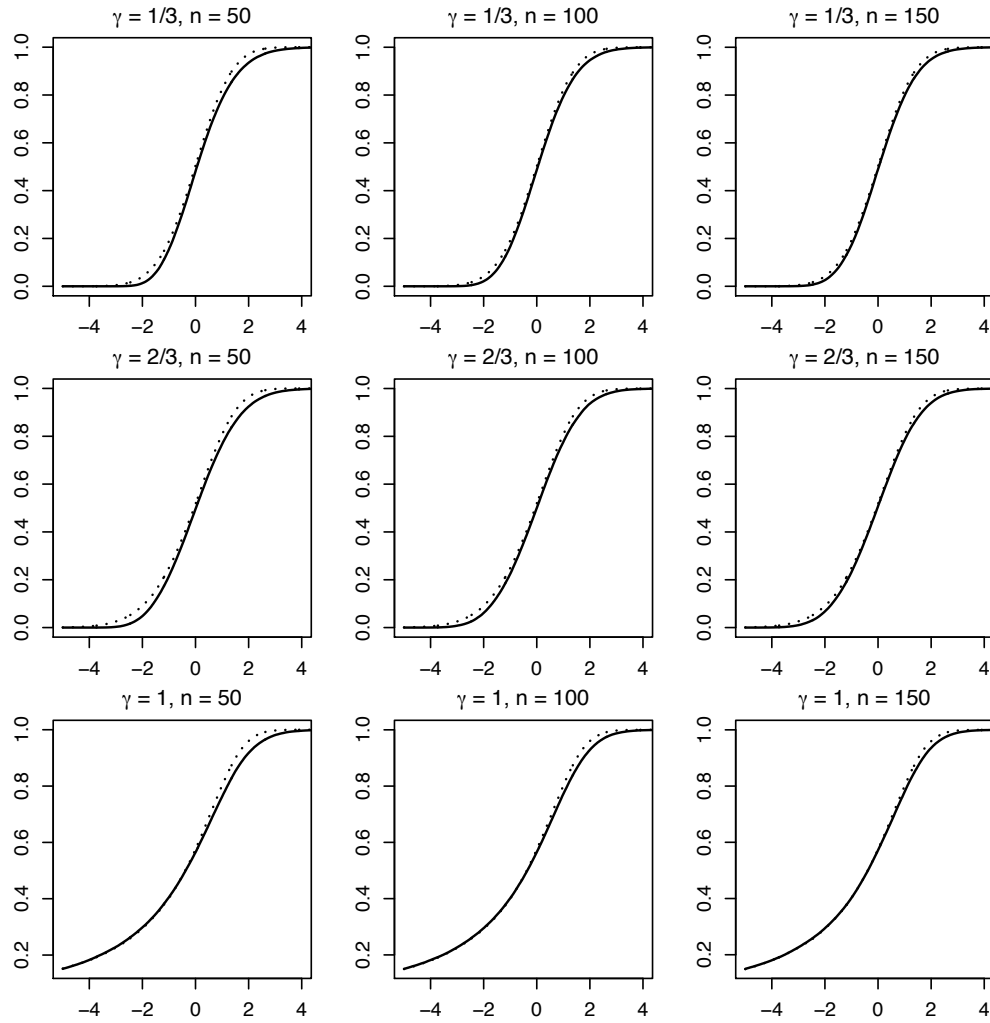


Figure 3.2: In solid black, distribution function of $\hat{u}$ for different values of γ and n . In dashed black, the distribution function of the approximation provided by the theorem.

3.3 Appendix

3.3.1 Proof of Lemma 1

We begin by mentioning some basic properties of the functions $h_i(x) = (\partial/\partial x^i) \log \Phi(x)$ for $i = 1, 2$ which we will use in this proof and in subsequent analyses. Suppose that Z follows a $N(x, 1)$ distribution truncated to the interval $(0, \infty)$. We saw that its cumulant generating function is $(x+t)^2/2 + \log \Phi(x+t) - x^2/2 + \log \Phi(x)$. Therefore, $\mathbb{E}_x[Z] = x + h_1(x)$ and $\text{Var}_x(Z) = 1 + h_2(x)$. [Mailhot \(1985\)](#) shows that $\text{Var}_x(Z)$, and hence $h_2(x)$, is strictly increasing. Furthermore, since $\text{Var}_x(Z)$ tends to 0 and 1 as x approaches $-\infty$ and ∞ , respectively, $h_2(x)$ increases from -1 to 0. As for $h_1(x)$, since $h_2(x)$ is negative everywhere, it is strictly decreasing. Moreover, as $x \rightarrow \infty$, $h_1(x)$ clearly approaches 0, and as $x \rightarrow -\infty$, it tends to $-\infty$ and behaves as $-x$ asymptotically. Indeed, by l'Hôpital's rule,

$$\lim_{x \rightarrow -\infty} \frac{h_1(x)}{x} = \lim_{x \rightarrow -\infty} h_2(x) = -1.$$

Now, take $\gamma = 1$ and assume that $\theta_n < t_n - \varepsilon$ for all large enough n and some $\varepsilon > 0$. Inconsistency of $\tilde{\theta}_n = Y_n$ is trivial, since, under selection $Y_n > t_n$. Furthermore, for all $x > 0$,

$$\mathbb{P}_{\theta_n}(Y_n > t_n + x \mid S_n) = \frac{\mathbb{P}_{\theta_n}(Y_n > t_n + x)}{\mathbb{P}_{\theta_n}(Y_n > t_n)} = \frac{\Phi\{n^{1/2}(\theta_n - t_n) - n^{1/2}x\}}{\Phi\{n^{1/2}(\theta_n - t_n)\}} \rightarrow 0$$

as $n \rightarrow \infty$, which implies that $Y_n - t_n \xrightarrow{d} 0$, where $\xrightarrow{d}$ denotes convergence in distribution. For the MLE, inverting the score equation gives

$$\begin{aligned} \mathbb{P}_{\theta_n}(\hat{\theta}_n \leq t_n - x \mid S_n) &= \mathbb{P}_{\theta_n}\{Y_n \leq t_n - x + n^{-1/2}h_1(-n^{1/2}x) \mid S_n\} \\ &= \mathbb{P}_{\theta_n}\{n(Y_n - t_n) \leq -nx + n^{1/2}h_1(-n^{1/2}x) \mid S_n\}. \end{aligned}$$

We are going to use that, as $n \rightarrow \infty$,

$$\begin{aligned} n \frac{Y_n - t_n}{t_n - \theta_n} \mid S_n &\xrightarrow{d} \text{Exp}(1); \\ -nx + n^{1/2}h_1(-n^{1/2}x) &\rightarrow x^{-1}; \end{aligned}$$

for all $x > 0$. For the first claim, let $z_n = n^{1/2}(t_n - \theta_n)$, which diverges to ∞ . l'Hôpital's

rule gives, for all $x > 0$,

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \mathbb{P}_{\theta_n} \{n(t_n - \theta_n)(Y_n - t_n) > x \mid S_n\} \\
&= \lim_{n \rightarrow \infty} \frac{\mathbb{P}_{\theta_n} \{n(t_n - \theta_n)(Y_n - t_n) > x\}}{\mathbb{P}_{\theta_n}(Y_n > t_n)} \\
&= \lim_{n \rightarrow \infty} \frac{\Phi(-z_n - xz_n^{-1})}{\Phi(-z_n)} \\
&= \lim_{n \rightarrow \infty} \left(1 - \frac{x}{z_n^2}\right) \exp\left(-\frac{x^2}{2z_n} - x\right) \\
&= \exp(-x).
\end{aligned}$$

For the second claim, let $z = -n^{1/2}x$ and observe that

$$\lim_{z \rightarrow -\infty} -z^2 - zh_1(z) = \lim_{z \rightarrow -\infty} \frac{z}{h_1(z)} h_2(z) = 1,$$

where we have used that $h_2(z) = -h_1(z)\{z + h_1(z)\}$ and that both $h_2(z)$ and $h_1(z)/z$ tend to -1 as $z \rightarrow -\infty$. From these results, it follows that $\mathbb{P}_{\theta_n}(\hat{\theta}_n \leq t_n - x \mid S_n) = \mathbb{P}\{\text{Exp}(1) \leq x^{-1}(t_n - \theta_n)^{-1}\} + o(1)$ for all $x > 0$ as $n \rightarrow \infty$. Since $(t_n - \theta_n)^{-1} < \varepsilon^{-1}$, the dominant term of the asymptotic expansion is bounded away from zero.

Now assume that $\gamma = 1$ and $\varphi_n(\theta_n) > \varepsilon > 0$ for all n . For the MOE we have that

$$\mathbb{P}_{\theta_n} \left\{ n^{1/2} |Y_n - \theta_n| > x \mid S_n \right\} \leq \varepsilon^{-1} \mathbb{P}_{\theta_n} \left\{ n^{1/2} |Y_n - \theta_n| > x \right\},$$

which can be made arbitrarily small by increasing x , showing the claim. The MLE satisfies the equation $Y_n = g_n(\hat{\theta}_n)$, where $g_n(x) = x + n^{-1/2}h_1\{n^{1/2}(x - t_n)\}$. Thus, for every $M > 0$, inverting $Y = g_n(\hat{\theta}_n)$ gives

$$\begin{aligned}
& \mathbb{P}_{\theta_n} \left\{ n^{1/2}(\hat{\theta}_n - \theta_n) \leq -M \mid S_n \right\} \\
&= \mathbb{P}_{\theta_n} \left\{ Y_n \leq g_n(\theta_n - n^{-1/2}M) \mid S_n \right\} \\
&\leq \mathbb{P}_{\theta_n} \left\{ Y_n \leq g_n(\theta_n) - n^{-1/2}Mg'_n(\theta_n - n^{-1/2}M) \mid S_n \right\} \\
&= \mathbb{P}_{\theta_n} \left\{ n^{1/2}(Y_n - g_n(\theta_n)) \leq -Mg'_n(\theta_n - n^{-1/2}M) \mid S_n \right\}.
\end{aligned}$$

Now, $n^{1/2}\{g_n(\theta_n) - \theta_n\} = h_1\{n^{1/2}(\theta_n - t_n)\}$. By assumption, $n^{1/2}(\theta_n - t_n)$ is bounded from below as n increases. Moreover, $h_1(x)$ is positive and decreasing, so $g_n(\theta_n) = \theta_n + O(n^{-1/2})$. Since $Y_n = \theta_n + O_p(n^{-1/2})$, we also have that $Y_n = g(\theta_n) + O_p(n^{-1/2})$. Furthermore, $g'_n(\theta_n - n^{-1/2}M) = 1 + h_2\{n^{1/2}(\theta_n - t_n) - M\}$ is bounded away from zero, as $1 + h_2(x)$ is bounded away from zero in sets that are bounded from below. Putting all together, we have that $\mathbb{P}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n) \leq -M \mid S_n\}$ can be made arbitrarily small uniformly in n by increasing M . A similar argument applied to $\mathbb{P}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n) \geq M \mid S_n\}$ shows that

$$\hat{\theta}_n = \theta_n + O_p(n^{-1/2}).$$

Finally, let $\gamma \in (0, 1)$ be fixed. We have seen that the selective model can be represented as a model involving two independent observations $Y_{n1} \sim N(\theta_n, (\gamma n)^{-1})$ and $Y_{n2} \sim N(\theta_n, \{(1 - \gamma)n\}^{-1})$, with selection $Y_{n1} > t_n$. Since $\tilde{\theta}_n$ has smallest variance among the unbiased estimators of θ_n , its variance is smaller than that of Y_{2n} , which is simply $\{(1 - \gamma)n\}^{-1}$. This shows the uniform consistency of $\tilde{\theta}_n$. For the MLE write, as before, $Y_n = g(\tilde{\theta}_n)$, with $g_n(x) = x + (\gamma/n)^{1/2}h_1\{n_1^{1/2}(x - t_n)\}$. A first-order Taylor expansion of g_n^{-1} around $g_n(\theta_n)$ gives

$$\hat{\theta}_n = \theta_n + \frac{Y_n - g_n(\theta_n)}{1 + \gamma h_2(Z_n)},$$

for some $Z_n \in \mathbb{R}$. Since $h_2(x) > -1$ for all x , the denominator of the last term is always greater than $1 - \gamma > 0$. Also, since $g_n(\theta_n) = \mathbb{E}_{\theta_n}[Y \mid S_n]$, we have that $|Y_n - g_n(\theta)| = O_p\{\text{Var}_{\theta_n}(Y \mid S_n)^{1/2}\} = O_p\{n^{-1/2}g'_n(\theta_n)\}$. Since $g'_n(x) = 1 + \gamma h_2\{n^{1/2}(x - t_n)\} < 1$, as $h_2(x)$ is always negative, $|Y_n - g_n(\theta)| = O_p(n^{-1/2})$ and the last claim follows.

3.3.2 Proof of Theorem 2

The main part of the proof concerns the validity of approximating the probability ratios based on the given model by probability ratios based on a normal distribution. We are going to write the proof for the case $\gamma = 1$ to avoid saturating the notation. For $\gamma < 1$ we would simply swap n by n_1 everywhere.

Let $X_n(u) = \log\{f(Y_1, \dots, Y_n; \theta_0 + n^{-1/2}u)/f(Y_1, \dots, Y_n; \theta_0)\}$ be the non-selective log-likelihood ratio around θ_0 . We can write

$$\frac{\varphi_n(\theta_0 + n^{-1/2}u)}{\varphi_n(\theta_0)} = \mathbb{E}_{\theta_0}[\exp\{X_n(u)\} \mid S_n] \equiv r_n(u),$$

where S_n is the selection event of the n -th model. A Taylor expansion gives

$$X_n(u) = ui(\theta_0)Z_n + \frac{u^2}{2n}l''_n\left(\theta_0 + \frac{\varepsilon}{n^{1/2}}\right), \quad |\varepsilon| \leq |u|,$$

where $Z_n = n^{-1/2}i(\theta_0)^{-1}l'_n(\theta_0)$. Similarly, the probability ratio based on the normal approximation can be written as

$$\frac{\tilde{\varphi}_n(u)}{\tilde{\varphi}_n(0)} = \mathbb{E}\left[\exp\left\{ui(\theta_0)Z - \frac{u^2}{2}i(\theta_0)\right\} \mid \tilde{S}_n\right] \equiv \tilde{r}_n(u),$$

where $Z \sim N(0, i(\theta_0)^{-1})$ and $\tilde{S}_n$ is the selection event induced on Z by the selection function $\tilde{p}_n(z; \theta_0)$. We therefore want to prove that $r_n(u) - \tilde{r}_n(u) \rightarrow 0$ uniformly in any

set of the form $[-M, M]$, $M > 0$. We break the proof in three steps.

Step 1: $\varphi_n(\theta_0)/\tilde{\varphi}_n(0) \rightarrow 1$ as $n \rightarrow \infty$.

Assumption 3 ensures that $Z_n \xrightarrow{d} Z$. Let F_n and F be the distribution functions of Z_n and Z , and write

$$\varphi_n(\theta_0) = \int_{\mathcal{P}_n} \tilde{p}_n(z; \theta_0) dF_n(z); \quad \tilde{\varphi}_n(0) = \int_{\mathcal{P}_n} \tilde{p}_n(z; \theta_0) dF(z).$$

We write the probabilities as Riemann–Stieltjes integrals to cover both the discrete and absolute continuous cases. Integration by parts gives

$$\varphi_n(\theta_0) - \tilde{\varphi}_n(0) = \left\{ F_n(z) - F(z) \right\} \tilde{p}_n(z; \theta_0) \Big|_{\partial \mathcal{P}_n} - \int_{\mathcal{P}_n} \{F_n(z) - F(z)\} \tilde{p}'_n(z; \theta_0) dz.$$

Since convergence in distribution of a one-dimensional sequence implies uniform convergence of the distribution functions, and since $\int_{\mathcal{P}_n} |\tilde{p}'_n(z; \theta_0)| dz = O(1)$, it follows that $\varphi_n(\theta_0) - \tilde{\varphi}_n(0) \rightarrow 0$ as $n \rightarrow \infty$. But $\varphi_n(\theta_0)$ is bounded away from zero, so we also have that $\varphi_n(\theta_0)/\tilde{\varphi}_n(0) \rightarrow 1$.

Step 2: $|\mathbb{E}_{\theta_0}[\exp(ui(\theta_0)Z_n)\tilde{p}_n(Z_n; \theta_0)] - \mathbb{E}[\exp(ui(\theta_0)Z)\tilde{p}_n(Z; \theta_0)]| \rightarrow 0$ uniformly in $[-M, M]$ for all $M > 0$.

To prove this we first show that, for all $M > 0$, the collection of random variables $\{\exp(ui(\theta_0)Z_n) : |u| \leq M, n \in \mathbb{N}\}$ is uniformly integrable under θ_0 . Fix any $u \in \mathbb{R}$ and let $m(x) = \mathbb{E}_{\theta_0}[\exp\{x l'_1(\theta_0)\}]$ for x in a ball around 0. We have that $\mathbb{E}_{\theta_0}[\exp(ui(\theta_0)Z_n)] = m(n^{-1/2}u)^n$. To show uniform integrability it suffices to prove that

$$\sup\{\mathbb{E}_{\theta_0}[\exp(ui(\theta_0)Z_n)^2] = m(n^{-1/2}2u)^n : |u| \leq M, n \in \mathbb{N}\} < \infty.$$

Using that $m(0) = 1$ and $m'(0) = 0$ (as the score has null expectation), a Taylor expansion of $\log m(x)$ gives

$$n \log m\left(\frac{2u}{n^{1/2}}\right) = 2u^2 \frac{\partial^2}{\partial x^2} \log m(\tilde{x}), \quad |\tilde{x}| \leq \frac{2|u|}{n^{1/2}},$$

so for some $K > 0$ and any large enough n ,

$$\sup_{|u| \leq M, n \in \mathbb{N}} m\left(\frac{2u}{n^{1/2}}\right)^n \leq \exp\left\{2M^2 \sup_{|x| \leq K} \frac{\partial^2}{\partial x^2} \log m(x)\right\} < \infty,$$

as $m(x)$ is infinitely differentiable. Thus, $\{\exp(ui(\theta_0)Z_n) : |u| \leq M, n \in \mathbb{N}\}$ is uniformly integrable, and, trivially, so is $\{\exp(ui(\theta_0)Z_n)\tilde{p}_n(Z_n; \theta_0) : |u| \leq M, n \in \mathbb{N}\}$.

This implies that for every $\varepsilon > 0$ there exists a constant $C_1 > 0$ such that

$$\sup_{|u| \leq M, n \in \mathbb{N}} \mathbb{E}_{\theta_0} [\exp(ui(\theta_0)Z_n) \tilde{p}_n(Z_n; \theta_0) \mathbf{1}\{\exp(ui(\theta_0)Z_n) \tilde{p}_n(Z_n; \theta_0) > C_1\}] < \varepsilon.$$

By a similar argument, there is a $C_2 > 0$ such that

$$\sup_{|u| \leq M, n \in \mathbb{N}} \mathbb{E} [\exp(ui(\theta_0)Z) \tilde{p}_n(Z; \theta_0) \mathbf{1}\{\exp(ui(\theta_0)Z) \tilde{p}_n(Z; \theta_0) > C_2\}] < \varepsilon.$$

Take $C = \max(C_1, C_2)$ and let $A_n(u) = \mathcal{P}_n \cap \{z: \exp(ui(\theta_0)z) \tilde{p}_n(z; \theta_0) \leq C\}$. As in step 1, write

$$\begin{aligned} & \int_{A_n(u)} \exp(ui(\theta_0)z) \tilde{p}_n(z; \theta_0) dF_n(z) - \int_{A_n(u)} \exp(ui(\theta_0)z) \tilde{p}_n(z; \theta_0) dF(z) \\ &= \left. \{F_n(z) - F(z)\} \tilde{p}_n(z; \theta_0) \exp(ui(\theta_0)z) \right|_{\partial A_n(u)} \\ & \quad - \int_{A_n(u)} \{F_n(z) - F(z)\} \tilde{p}'_n(z; \theta_0) \exp(ui(\theta_0)z) dz \\ & \quad - \int_{A_n(u)} \{F_n(z) - F(z)\} ui(\theta_0) \tilde{p}_n(z; \theta_0) \exp(ui(\theta_0)z) dz. \end{aligned}$$

By definition of $A_n(u)$ the first and third terms go to zero uniformly for $u \in [-M, M]$. For the second term, consider any bounded sequence $(u_n)_{n \in \mathbb{N}} \subseteq [-M, M]$. The sequence of functions $h_n(z) = \{F_n(z) - F(z)\} \tilde{p}'_n(z; \theta_0) \exp(u_n i(\theta_0)z) \mathbf{1}\{z \in A_n(u_n)\}$ converges pointwise to zero and has

$$|h_n(z)| \leq |[1 - F(z) - \{1 - F_n(z)\}] \exp(u_n i(\theta_0)z) \tilde{p}'_n(z; \theta_0)| \leq 2S(z) \exp\{Mi(\theta_0)z\} |\tilde{p}'_n(z; \theta_0)|,$$

where $S(x)$ is defined, for a fixed x , as the least upper bound of the sequence $\{1 - F_n(x): n \in \mathbb{N}\}$. By Theorem 2 of [Kozakiewicz \(1947\)](#), $S(z) \exp\{Mi(\theta_0)z\}$ is integrable in $\mathbb{R}$. Therefore, by the boundedness assumption on $\tilde{p}'_n(z; \theta_0)$, the second term vanishes uniformly in compact sets of u by the Dominated Convergence Theorem. This verifies the claim, as the integrals outside $A_n(u)$ are bounded by 2ε by definition, which can be made arbitrarily small by increasing C .

Step 3: $\mathbb{E}_{\theta_0} [\exp\{X_n(u)\} - \exp\{ui(\theta_0)Z_n - \frac{u^2}{2}i(\theta_0)\}] p_n(Y_1, \dots, Y_n) \rightarrow 0$ uniformly in $[-M, M]$ for all $M > 0$.

Applying the Mean Value Theorem and the Cauchy-Schwarz inequality gives the following

bound:

$$\begin{aligned}
& \mathbb{E}_{\theta_0} \left[\left| \exp \{X_n(u)\} - \exp \left\{ ui(\theta_0)Z_n - \frac{u^2}{2}i(\theta_0) \right\} \right| p_n(Y_1, \dots, Y_n) \right] \\
&= \frac{u^2}{2} \mathbb{E}_{\theta_0} \left[\left| \frac{1}{n} l_n'' \left(\theta_0 + \frac{\varepsilon}{n^{1/2}} \right) + i(\theta_0) \right| \exp \left\{ ui(\theta_0)Z_n + \frac{u^2}{2}\varepsilon_2 \right\} \right] \\
&\leq \frac{u^2}{2} \mathbb{E}_{\theta_0} \left[\left\{ \frac{1}{n} l_n'' \left(\theta_0 + \frac{\varepsilon}{n^{1/2}} \right) + i(\theta_0) \right\}^2 \right]^{1/2} \mathbb{E}_{\theta_0} [\exp \{4ui(\theta_0)Z_n\}]^{1/4} \mathbb{E}_{\theta_0} [\exp \{2u^2\varepsilon_2\}]^{1/4}
\end{aligned}$$

for some $|\varepsilon_2| \leq |n^{-1}l_n''(\theta_0 + n^{-1/2}\varepsilon) + i(\theta_0)|$. The third factor is uniformly bounded by the analysis of step 2. We also have

$$n^{-1}l_n''(\theta_0 + n^{-1/2}\varepsilon) + i(\theta_0) = n^{-1}l_n''(\theta_0) + i(\theta_0) + n^{-3/2}\varepsilon l_n'''(\theta_0 + n^{-1/2}\varepsilon_3), \quad |\varepsilon_3| \leq |\varepsilon| \leq |u|,$$

so

$$\begin{aligned}
\left\{ \frac{1}{n} l_n'' \left(\theta_0 + \frac{\varepsilon}{n^{1/2}} \right) + i(\theta_0) \right\}^2 &= \left\{ \frac{1}{n} l_n''(\theta_0) + i(\theta_0) \right\}^2 \\
&+ 2 \left\{ \frac{1}{n} l_n''(\theta_0) + i(\theta_0) \right\} n^{-3/2} \varepsilon l_n'''(\theta_0 + n^{-1/2}\varepsilon_3) \\
&+ \left\{ n^{-3/2} \varepsilon l_n'''(\theta_0 + n^{-1/2}\varepsilon_3) \right\}^2.
\end{aligned}$$

The expectation of the first term is the variance of $l_n''(\theta_0)$, which is equal to $n^{-1}\text{Var}_{\theta_0}\{l_1''(\theta_0)\} \rightarrow 0$, and the second and third terms vanish uniformly by assumption 5. Finally,

$$\mathbb{E}_{\theta_0} [\exp \{2u^2\varepsilon_2\}] \leq \mathbb{E}_{\theta_0} \left[\exp \left\{ 2u^2 |n^{-1}l_n''(\theta_0 + n^{-1/2}\varepsilon) + i(\theta_0)| \right\} \right],$$

which is also uniformly bounded by assumption 4.

Step 4: $r_n(u) - \tilde{r}_n(u) \rightarrow 0$ uniformly in $[-M, M]$ for all $M > 0$.

Writing

$$\begin{aligned}
& \varphi_n(\theta_0) \{r_n(u) - \tilde{r}_n(u)\} \\
&= \mathbb{E}_{\theta_0} [\exp \{X_n(u)\} p_n(Y_1, \dots, Y_n)] - \frac{\varphi_n(\theta_0)}{\tilde{\varphi}_n(0)} \mathbb{E}_{\theta_0} [\exp \{ui(\theta_0)Z - u^2i(\theta_0)/2\} \tilde{p}_n(Z; \theta_0)]
\end{aligned}$$

and applying steps 1-3 gives the result.

To complete the proof of the theorem, let $Z_{n_1} = n_1^{-1/2}i(\theta_0)^{-1}l'_{n_1}$ and $Z_{n_2} = n_2^{-1/2}i(\theta_0)^{-1}l'_{n_2}$. Since Z_{n_2} is unaffected by selection, it converges in distribution to a $N(0, i(\theta_0)^{-1})$ under selection. To prove the claim about the asymptotic distribution of Z_{n_1} , we need to show

that $\mathbb{P}_{\theta_0}(Z_{n_1} \leq x \mid S_{n_1}) - G_{n_1}(x)$ converges uniformly to zero, which follows from the same argument used to prove step 1.

Finally, for any converging sequence $u_n \rightarrow u$, a Taylor expansion gives

$$X_n(u_n) = ui(\theta_0)\{\gamma^{1/2}Z_{n_1} + (1 - \gamma^{1/2})Z_{n_2}\} - \frac{u^2}{2}i''(\theta_0) + o_p(1)$$

where the $o_p(1)$ error holds under the non-selective distribution. However, since the selection probability at the true parameter is assumed to be bounded away from zero, the expansion also holds conditionally on selection. To see this, simply note that for every positive random variable X and selection event S , we have $\mathbb{P}(X > \varepsilon \mid S) \leq \mathbb{P}(X > \varepsilon)/\mathbb{P}(S)$, which means that orders of magnitude are preserved by conditioning on events with non-vanishing probability.

Chapter 4

Non-informative priors for selected parameters

We propose two types of non-informative prior densities for selection models that lead to well-calibrated posterior inferences under repeated sampling. By analogy with the derivation of probability-matching priors without selection, we first restrict the problem to the class of normal-location models, and then extend the proposal to more general settings by appealing to asymptotic considerations.

4.1 Probability matching in normal models

Non-informative priors allow the derivation of posterior distributions without explicit incorporation of prior information. They serve different purposes: the resulting posterior can be employed as a reference against which posteriors derived from subjective priors can be compared; they can be used to derive frequentist methods that retain some of the appealing characteristics of Bayesian methods; and they allow to carry out an analysis when very little or no prior information is available. A standard requirement on these priors is that they are approximately well-calibrated under repeated sampling from a fixed parameter. More precisely, they are required to satisfy

$$\mathbb{P}_\theta\{\psi \leq \Pi^{-1}(\alpha|Y)\} = \alpha + \varepsilon(\alpha, \theta), \quad (4.1)$$

where $\Pi(\psi|Y)$ is the marginal posterior distribution function of the interest parameter and $\varepsilon(\alpha, \theta)$ is small for all $\alpha \in (0, 1)$ and $\theta \in \Theta$. Condition (4.1) ensures that posterior claims about the parameter have approximate validity in a frequentist sense. For example, that

$(1 - \alpha)$ -credible intervals for ψ are also approximate $(1 - \alpha)$ -confidence intervals. In the context of selective inference, this condition would be required to hold with respect to the relevant selective distribution, as per Section 2.1.

In non-selective regimes involving independent and identically distributed observations from a regular model, the quantile-matching requirement holds with an error decreasing at rate $n^{-1/2}$ for any prior that places positive probability around the true parameter, where n is the sample size. In addition, certain priors, known as probability-matching priors, improve on this error rate, lowering it to n^{-1} in general, and to $n^{-3/2}$ for some distributions; see [Datta and Mukerjee \(2004\)](#). Unfortunately, the formal analysis leading to the derivation of these priors cannot be replicated in selection models, as it relies on the asymptotic normality on the posterior, which does not hold under selection except in trivial cases.

[Yekutieli \(2012\)](#) argues that non-informative priors should not be altered in the presence of selection. In particular, for a selective normal-location model, the improper uniform prior $\pi(\theta) \propto 1$ is suggested. This prior is also used by [Panigrahi and Taylor \(2018\)](#) for the case where θ is a linear regression parameter. As explained in Chapter 1, we sustain that a non-informative prior *should* be appropriately modified in the presence of selection, to account for the change of sampling distribution. As the following result illustrates, not doing so can lead to posterior inferences with poor frequentist properties.

Proposition 3. Let $Y \sim N(\theta, \sigma^2)$, with $\sigma^2 > 0$ known, and $p(y) = \mathbf{1}(y > t)$ for some fixed $t \in \mathbb{R}$, and let

$$\Pi(\theta|y) = \frac{\int_{-\infty}^{\theta} \phi\{\sigma^{-1}(\tilde{\theta} - y)\} \varphi(\tilde{\theta})^{-1} d\tilde{\theta}}{\int_{-\infty}^{\infty} \phi\{\sigma^{-1}(\tilde{\theta} - y)\} \varphi(\tilde{\theta})^{-1} d\tilde{\theta}}$$

be the selective posterior distribution function based on the uniform prior $\pi(\theta) \propto 1$. Then,

$$\mathbb{P}_{\theta}\{\theta \leq \Pi^{-1}(\alpha|Y)|S\} < \alpha \quad \forall(\alpha, \theta) \in (0, 1) \times \mathbb{R}.$$

The first part of Proposition 3 says that, when selection takes the form of one-sided truncation, the uniform prior produces inferences that overstate values of θ with low selection probability. So, in a way, correcting for selection overcompensates the selection bias. In addition, simulation results at the end of this section suggest that this claim holds for any selection mechanism with a increasing selection function, such as $p(y) = \mathbb{P}(y + W > t)$ for some randomisation noise W , although we were not able to prove this. In Section 4.2 we will see that this overcompensation also occurs in non-normal models under standard non-informative priors. The second part of the proposition says that credible intervals derived from these posteriors have infinite expected length when considered under repeated sampling from a fixed parameter. This is not surprising, as

in this case the model is highly misspecified. Recall that, for example, the MLE does not have a mean in this model. This suggests a need for regularisation in regions of the parameter space with low selection probability.

Unlike in the non-selective case, where the uniform prior leads to exact probability matching, it can be shown that in a selective normal-location model exact matching can only be achieved with a data-dependent prior. Before doing so, let us introduce the concept of a confidence distribution. Consider the model $Y \sim N(\theta, \sigma^2)$, with σ^2 known, and an arbitrary selection function $p(y)$ which does not vanish almost everywhere. Conditionally on selection, the p -value function

$$H(\theta; y) = \mathbb{P}_\theta(Y \geq y | S)$$

is uniformly distributed over $(0, 1)$ when θ is the true parameter value. Furthermore, if y is such that $0 < H(0; y) < 1$, then $\theta \rightarrow H(\theta; y)$ is a distribution function. To see this, assume that $\sigma^2 = 1$ for simplicity and write

$$H(\theta; y) = \frac{\mathbb{P}(S, Y \geq y)}{\varphi(\theta)} = \frac{\int_y^\infty \phi(\tilde{y} - \theta)p(\tilde{y})d\tilde{y}}{\int_{-\infty}^\infty \phi(\tilde{y} - \theta)p(\tilde{y})d\tilde{y}}.$$

Differentiating with respect to θ gives

$$\frac{\partial}{\partial \theta} H(\theta; y) = \frac{\mathbb{P}_\theta(S, Y \geq y)}{\varphi(\theta)} \{ \mathbb{E}_\theta[Y | S, Y \geq y] - \mathbb{E}_\theta[Y | S] \} > 0$$

for all θ . Furthermore, $0 < H(0; y)$ implies that $\int_y^\infty \phi(\tilde{y} - \theta)p(\tilde{y})d\tilde{y} \geq K\{\Phi(y_2 - \theta) - \Phi(y_1 - \theta)\}$ for some $y < y_1 < y_2$ and $K > 0$, and we clearly have that $\mathbb{P}(S, Y < y) \leq \Phi(y - \theta)$. Therefore

$$\frac{1}{H(\theta; y)} = \frac{\varphi(\theta)}{\mathbb{P}(S, Y \geq y)} = 1 + \frac{\mathbb{P}(S, Y < y)}{\mathbb{P}(S, Y \geq y)} \leq 1 + \frac{1}{K} \frac{\Phi(y - \theta)}{\Phi(y_2 - \theta) - \Phi(y_1 - \theta)} \rightarrow 1$$

as $\theta \rightarrow \infty$. An analogous argument, using that $H(0; y) < 1$, shows that the limit as $\theta \rightarrow -\infty$ is zero, so $H(\theta; y)$ is indeed a distribution function.

Considered as a random function of Y , $H(\theta; Y)$ is an example of a so-called confidence distribution. In a one-dimensional statistical model $\{F(y; \theta) : \theta \in \Theta \subseteq \mathbb{R}\}$, a confidence distribution is any function $G(\theta; Y)$ that satisfies the following two conditions:

1. $G(\theta; Y)$ is uniformly distributed over $(0, 1)$ when θ is the true parameter;
2. $\theta \rightarrow G(\theta; y)$ is a distribution function for every y .

Note that, in the case discussed here, $H(\theta; y)$ is a distribution function if $0 < H(\theta; y) < 1$

for some θ , but since $H(\theta; Y) \sim U(0, 1)$ under θ , it follows that $\mathbb{P}_\theta\{H(\theta; Y) \in (0, 1) \mid S\} = 1$, so the second condition is satisfied almost surely.

Confidence distributions can be thought of as probability distributions over the parameter space that are induced by the data and can be derived without prior information. They are, in essence, a modern formulation of Fisher's fiducial probability (Fisher, 1930), and have attracted much attention in recent years due to their joint frequentist-Bayesian spirit; see Xie and Singh (2013) for a recent review. Confidence distributions can be used to derive confidence intervals with any desired level since, by definition, the interval $[H^{-1}(\alpha_1; Y), H^{-1}(\alpha_2; Y)]$ is a confidence interval with coverage $\alpha_2 - \alpha_1$ for any $0 < \alpha_1 < \alpha_2 < 1$. In addition, $H^{-1}(1/2; Y)$ is a median-unbiased estimator of θ , as it satisfies $\mathbb{P}_\theta\{H^{-1}(1/2; Y) \leq \theta \mid S\} = 1/2$.

Given a confidence distribution $H(\theta; y)$, one can trivially construct an exact probability-matching prior by requiring that the posterior distribution function $\Pi(\theta|y)$ equals $H(\theta; y)$. This type of construction appears, for example, in Fraser et al. (2010), and can be used as a basis for deriving probability-matching priors in more complex scenarios, as we will do in the next section. The resulting prior density is usually data-dependent, and is computed as

$$\pi_y(\theta) \propto -\frac{\frac{\partial}{\partial \theta} H(\theta; y)}{\frac{\partial}{\partial y} H(\theta; y)}.$$

In the case of the selective normal-location model, this is, in addition, the only *sensible* prior that provides exact matching. Indeed, consider the matching equation

$$\mathbb{P}_\theta\{\Pi(\theta|Y) \leq \alpha \mid S\} = \alpha, \quad \alpha \in (0, 1).$$

Since the model is stochastically increasing in θ , $\Pi(\theta|y)$ ought to be a decreasing function of y for every θ . Denoting its inverse with respect to y by $l_\theta(\alpha)$, we have that

$$\mathbb{P}_\theta\{Y \geq l_\theta(\alpha) \mid S\} = H\{\theta; l_\theta(\alpha)\} = \alpha, \quad \alpha \in (0, 1).$$

Letting $\alpha = \Pi(\theta|y)$ gives the equality.

For a general selection mechanism, the functional form of $\pi_y(\theta)$ does not admit a simple expression. For the model $Y \sim N(\theta, n^{-1})$ with selection $Y > t$, it can be expressed as

$$\pi_y(\theta) \propto 1 - \frac{h_1\{n^{1/2}(\theta - t)\}}{h_1\{n^{1/2}(\theta - y)\}}$$

where $h_1(x) = (\partial/\partial x) \log \Phi(x) = \phi(x)/\Phi(x)$. In the data splitting case, where $Y = \gamma Y_1 + (1 - \gamma) Y_2 \sim N(\theta, n^{-1})$ for $Y_1 \sim N(\theta, n_1^{-1})$ and $Y_2 \sim N(\theta, n_2^{-1})$ independently,

$\gamma = n_1/n$, and selection is $Y_1 > t$, it is given by

$$\begin{aligned} \pi_y(\theta) \propto & \gamma^{1/2} \frac{\phi(n_1^{1/2}(\theta - t))}{\phi(n_1^{1/2}(\theta - y))} \\ & \times \left\{ 1 - \Phi \left\{ \frac{n^{1/2}(y - \theta + \gamma(\theta - t))}{(1 - \gamma)^{1/2}} \right\} - \frac{\int_y^\infty n^{1/2} \phi(n^{1/2}(\tilde{y} - \theta)) p(\tilde{y}) d\tilde{y}}{\Phi(n_1^{1/2}(\theta - t))} \right\} \\ & + \Phi \left\{ \frac{n_1^{1/2}(y - t)}{(1 - \gamma)^{1/2}} \right\}. \end{aligned} \quad (4.2)$$

The derivation of this expression can be found in the appendix. Since evaluation of $\pi_y(\theta)$ requires computing integrals and dividing by small numbers, it can sometimes be numerically unstable. As a computationally lighter alternative, we will also consider the Jeffreys prior of the selective model, given by

$$\pi_J(\theta) \propto \left\{ 1 + \frac{\partial^2}{\partial \theta^2} \log \varphi(\theta) \right\}^{1/2}.$$

For the models considered before this prior can be written as

$$\pi_J(\theta) \propto \left[1 + \gamma h_2 \{ n_1^{1/2}(\theta - t) \} \right]^{1/2},$$

where $h_2(x) = (\partial^2 / \partial x^2) \log \Phi(x) = -x h_1(x) - h_1(x)^2$.

Note that both priors depend on the sample size, which is unusual in non-selective settings. In Figure 4.1 we have plotted both priors and the resulting posterior densities for this model, with $n = 20$, $\gamma = 1$, $t = 0$, and $y = 0.2$. The uniform prior and its posterior density is also plotted for comparison. The posterior densities corresponding to the two proposed priors are virtually identical, a behaviour that extends to other choices of parameters and observations, which suggests that Jeffreys prior is well-calibrated from a frequentist viewpoint. Our attempts to prove this claim formally were unsuccessful, but we provide empirical evidence of it in Table 4.1. The table shows the coverage of the intervals $(-\infty, \Pi^{-1}(\alpha|Y)]$ for this model with $n = 20$, $t = 0$, and different combinations of (θ, γ, α) , computed via numerical integration. The posterior distributions were constructed using the uniform prior (U) and Jeffreys prior (J). We see that the latter option performs better in most cases, and is considerably superior to the uniform prior when all the data is used for selection (the case $\gamma = 1$).

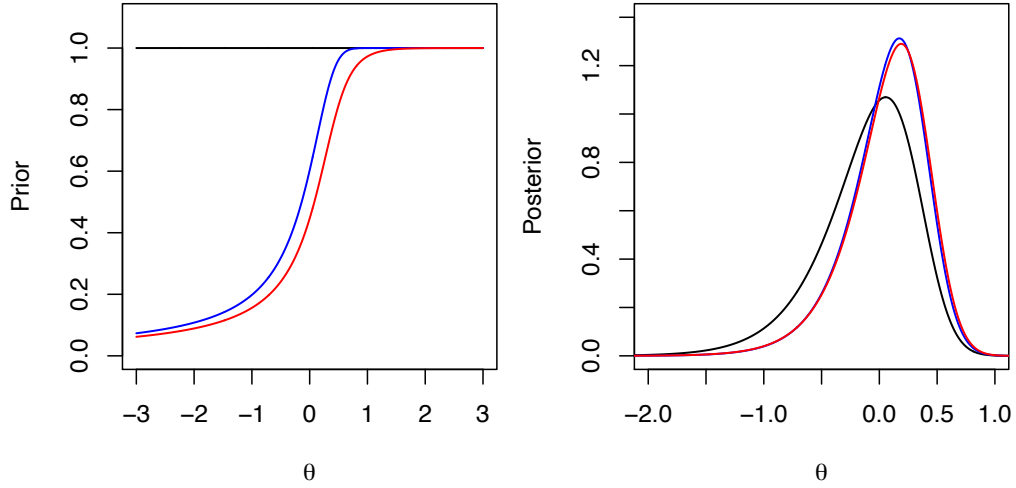


Figure 4.1: Left: uniform prior (black), probability-matching prior for $y = 0.2$ (red), and Jeffreys prior (blue) for the normal model. Right: the resulting posteriors for $y = 0.2$ (the red density overlaps the blue one in this panel).

α	0.05		0.1		0.25		0.5		0.75		0.9		0.95	
Prior	U	J	U	J	U	J	U	J	U	J	U	J	U	J
$\theta = -0.5$														
$\gamma = .5$	.045	.049	.092	.099	.237	.250	.487	.501	.742	.751	.896	.901	.948	.950
$\gamma = .75$	.035	.047	.075	.096	.209	.247	.458	.502	.724	.754	.889	.903	.944	.952
$\gamma = 1$	.020	.051	.041	.098	.110	.234	.250	.465	.448	.714	.637	.879	.738	.938
$\theta = 0$														
$\gamma = .5$	.043	.049	.088	.099	.228	.248	.473	.498	.730	.750	.890	.901	.945	.950
$\gamma = .75$	.038	.050	.077	.099	.199	.243	.427	.492	.693	.748	.872	.901	.935	.952
$\gamma = 1$	.033	.056	.065	.108	.161	.251	.331	.477	.537	.710	.709	.868	.794	.929
$\theta = 0.5$														
$\gamma = .5$	.048	.051	.095	.101	.238	.251	.479	.498	.728	.747	.886	.897	.941	.948
$\gamma = .75$	.048	.052	.095	.103	.236	.255	.471	.501	.713	.744	.871	.893	.930	.945
$\gamma = 1$	.048	.052	.096	.105	.237	.259	.469	.509	.702	.749	.852	.890	.908	.939

Table 4.1: Estimated coverages of $(-\infty, \Pi^{-1}(\alpha|Y)]$ for the normal-location model derived from the uniform (U) and Jeffreys (J) priors.

4.2 One-dimensional exponential family

The derivation of the probability-matching prior in the normal model is only useful to the extent that it can be used to devise priors for other models. If the model is normal we can provide inference directly via the confidence distribution. In this section we use the probability-matching and Jeffreys priors discussed in the previous section to derive non-informative priors for selected parameters of a one-dimensional exponential family, given a random sample from it. An attractive feature of exponential families that is useful here is that they always admit a sufficient statistic of the same dimension as the parameter. They are, in fact, the only regular models for which this holds, by the Pitman–Koopman–Darmois Theorem.

Suppose that we have a random sample $Y_1, \dots, Y_n$ from a full exponential family, with density or mass function

$$f(y_i; \theta) = h(y_i) \exp \{ \eta(\theta) s(y_i) - A(\theta) \}.$$

We assume that the MLE of the non-selective model, $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \eta(\theta) \sum_{i=1}^n s(y_i) - nA(\theta)$, exists for all realisations of the data (note that we are using the same notation for the MLE as in the previous chapter, even though here we are maximising the non-selective likelihood rather than the selective one). As in the previous chapter, let us assume that the sample is divided in two sets of sizes n_1 and n_2 , $\{Y_1, \dots, Y_{n_1}\}$ and $\{Y_{n_1+1}, \dots, Y_n\}$, with respective maximum likelihood estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, and that selection only uses the first set of samples, so that it occurs if $(Y_1, \dots, Y_{n_1})^T \in E_{n_1}$ for some event $E_{n_1} \subseteq \mathbb{R}^{n_1}$. Let us denote by $p_{\hat{\theta}_1}(\hat{\theta}_1) = \mathbb{P}(S \mid \hat{\theta}_1)$ the selection function of $\hat{\theta}_1$, which is independent of θ , as $\hat{\theta}_1$ is sufficient for θ .

Heuristically, a non-informative prior density for θ may be derived as follows. Let $i(\theta) = -\mathbb{E}_{\theta}[(\partial^2 / \partial \theta^2) \log f(Y_i; \theta)]$ be the Fisher information corresponding to a single observation. Any function $g(\theta)$ satisfying $g'(\theta) = i(\theta)^{1/2}$ is known as a variance-stabilising transformation, as it satisfies

$$n^{1/2} \{g(\hat{\theta}) - g(\theta)\} \xrightarrow{d} N(0, 1)$$

by the Delta Method. In particular, for $i = 1, 2$, we have that $n_i^{1/2} \{g(\hat{\theta}_i) - g(\theta)\} \xrightarrow{d} N(0, 1)$ as $n_i \rightarrow \infty$. Thus, for large values of n_1 and n_2 , the original selective model is approximately equivalent to a selective normal-location model involving two independent observations, $U_1 \sim N(g(\theta), n_1^{-1})$ and $U_2 \sim N(g(\theta), n_2^{-1})$, and selection function $p(u_1, u_2) = p_{\hat{\theta}_1} \{g^{-1}(u_1)\}$. This suggests choosing as a non-informative prior for the parameter $\nu = g(\theta)$ a prior that is reliable in the limiting normal model. Denoting the chosen prior by $\pi_{\nu}(\nu)$, the resulting prior in the original parametrisation is given by $\pi_{\theta}(\theta) \propto i(\theta)^{1/2} \pi_{\nu} \{g(\theta)\}$. In

view of the previous discussion and the simulation results, we will take $\pi_\psi(\psi)$ to be either the probability-matching prior or the Jeffreys prior. Note that, in the absence of selection, both choices give $\pi_\nu(\nu) \propto 1$, so $\pi_\theta(\theta)$ specialises to the Jeffreys prior of the non-selective model, $\pi_\theta(\theta) \propto i(\theta)^{1/2}$.

As an example, consider a model involving two binomial observations $Y_1 \sim \text{Bin}(n_1, \theta)$ and $Y_2 \sim \text{Bin}(n_2, \theta)$ and selection event $n_1^{-1}Y_1 \geq 0.5$, where $\theta \in [0, 1]$. The variance-stabilising transformation of this model is $g(\theta) = \sin^{-1}(\theta^{1/2})$. The left panel of Figure 4.2 shows the standard, non-selective Jeffreys prior $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$, and the two non-informative selective priors for $n_1 = 8$, $n_2 = 2$, $y_1 = 4$, and $y_2 = 1$. The right panel shows the corresponding posterior densities. As in the normal case, the two selective priors are almost identical and favour parameter values with large selection probability.

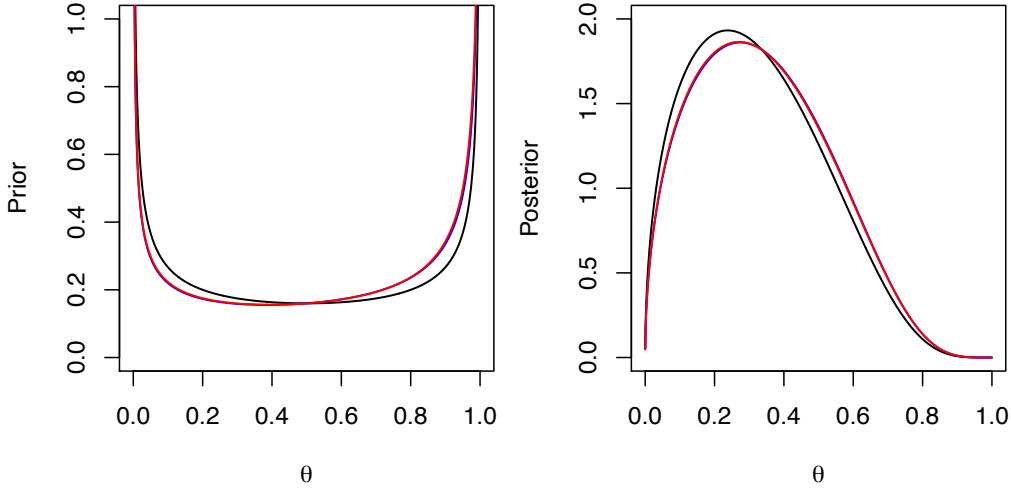


Figure 4.2: Left: non-selective Jeffreys prior (black), probability-matching prior for $y_1 = 4$ and $y_2 = 1$ (red), and Jeffreys prior (blue) for the binomial model. Right: the resulting posteriors (the red lines overlap the blue ones).

The following examples illustrate the repeated-sampling performance of the proposed priors.

Example 4.2.1. Exponential. Let $Y_1, \dots, Y_n$ be a random sample from an exponential distribution with rate parameter $\theta > 0$. For this model the variance-stabilising transformation is $g(\theta) = \log(\theta)$. We consider the selection event $\hat{\theta}_1 > 1$, where $\hat{\theta}_1$ is the maximum likelihood estimator of θ based on a subsample of size $n_1 = 0.8 \times n$. For the simulation we consider the values of $n = 10, 30, 80$, and for each sample size we consider three true parameter values defined to satisfy $\varphi_n(\theta_0) = 0.1, 0.5, 0.9$, corresponding to situations with small, moderate and large selection probabilities, respectively. For each pair (n, θ_0) , we plot the coverage of the interval $(-\infty, \Pi^{-1}(\alpha|Y_1, \dots, Y_n)]$ as a function of α for the non-

selective Jeffreys prior $\pi(\theta) \propto \theta^{-1}$ and for the two non-informative priors proposed in this work. The coverages were approximated via 10^4 simulations from the conditional model $(Y_1, \dots, Y_n)^T | \hat{\theta}_1 > 1$. The results can be found in Figure 4.3. The performances of the selective Jeffreys and probability-matching priors are practically identical and significantly better than that of the non-selective prior.

Example 4.2.2. Inverse Gaussian. Let $Y_1, \dots, Y_n$ be a random sample from an inverse Gaussian distribution with unknown mean $\theta > 0$ and known shape parameter $\lambda = 1$, with density

$$f(y_i; \theta) = \frac{1}{(2\pi y_i^3)^{1/2}} \exp \left\{ -\frac{(y_i - \theta)^2}{2\theta^2 y_i} \right\}.$$

The variance-stabilising transformation for this model can be easily seen to be $g(\theta) = -2\theta^{-1/2}$, and the non-selective Jeffreys prior $\pi(\theta) \propto \theta^{-3/2}$. We consider the same settings as before. The selection event is $\hat{\theta}_1 > 1$, where $\hat{\theta}_1$ is the maximum likelihood estimator of θ based on a subsample of size $n_1 = 0.8 \times n$ for $n = 10, 30, 80$, and we define the true values of the parameter in the same way as before. The results are similar to those of the exponential model.

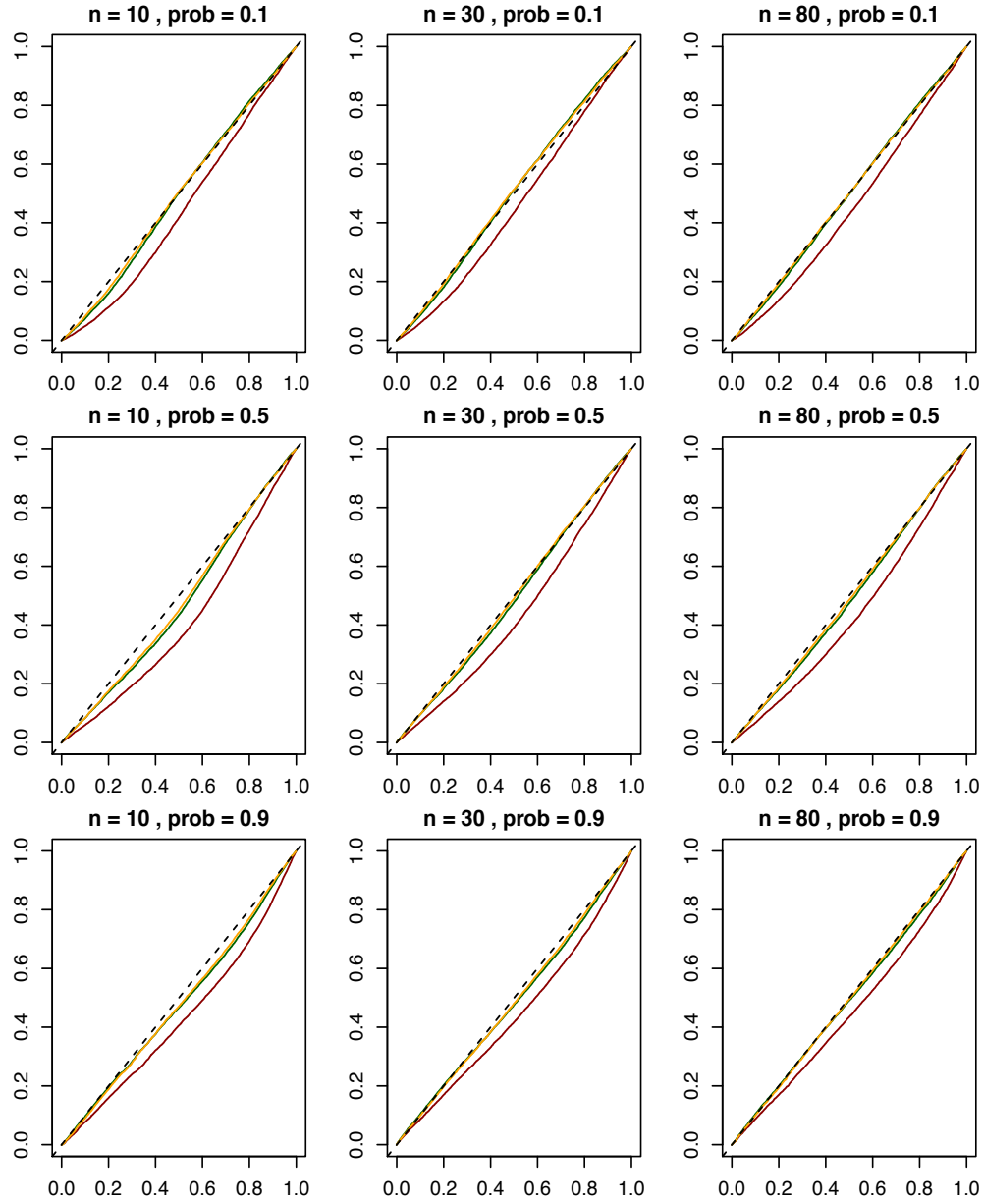


Figure 4.3: Example 4.2.1. Coverage of the interval $(-\infty, \Pi^{-1}(\alpha|Y_1, \dots, Y_n)]$ as a function of α for the non-selective Jeffreys prior (red), the probability-matching prior (orange), and the selective Jeffreys prior (green). The orange lines partially overlap the green ones.

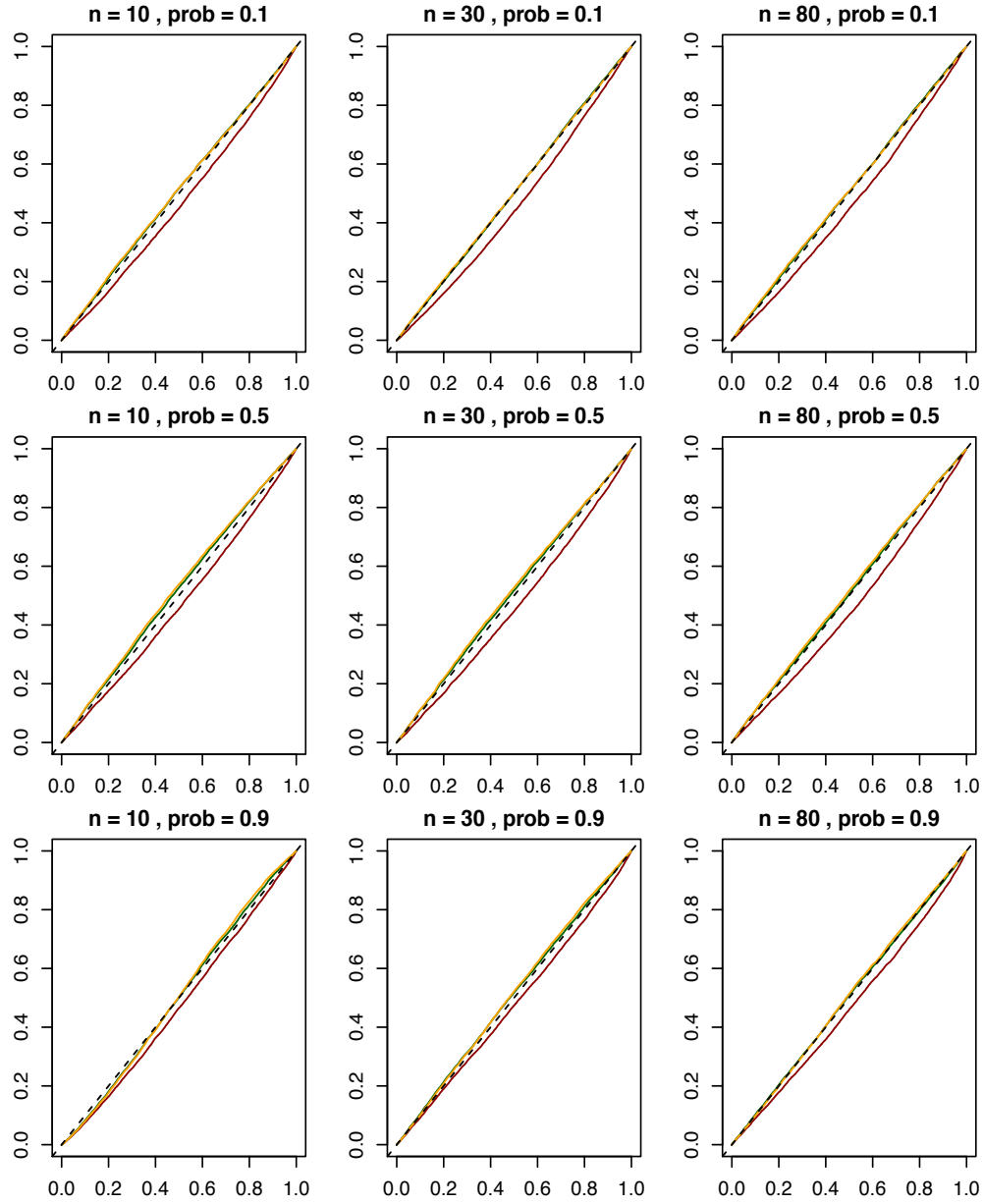


Figure 4.4: Example 4.2.2. Coverage of the interval $(-\infty, \Pi^{-1}(\alpha|Y_1, \dots, Y_n)]$ as a function of α for the non-selective Jeffreys prior (red), the probability-matching prior (orange), and the selective Jeffreys prior (green). The orange lines partially overlap the green ones.

4.3 Non-informative priors in the presence of nuisance parameters

We now consider situations where the full parameter θ is multidimensional and the selected parameter, ψ , is one-dimensional. Following similar arguments to the previous section, we provide an heuristic derivation of non-informative priors for ψ . Again, we assume that the non-selective model is a full exponential family, with density or mass function

$$f(y_i; \theta) = h(y_i) \exp \{ \eta(\theta)^T s(y_i) - A(\theta) \}, \quad \theta \in \Theta \subseteq \mathbb{R}^p,$$

and, as before, we assume that selection is based on a random sample of size n_1 , $Y_1, \dots, Y_{n_1}$, and that there is a second set of samples of size n_2 , $Y_{n_1+1}, \dots, Y_n$, available for inference.

Suppose that we can write $\theta = (\psi, \chi)$, and let us reparametrise the model to (ψ, λ) , where $\lambda \equiv \lambda(\psi, \chi)$ is orthogonal to ψ as per the definition of [Cox and Reid \(1987\)](#). This means that the Fisher information matrix in the latter parametrisation has zeroes in all the non-diagonal entries of the first row and column, or equivalently, that the maximum likelihood estimators $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent, in the absence of selection. Such parametrisation always exists when ψ is one-dimensional, and admits a very simple expression in terms of the cumulant generating function $A(\theta)$ when ψ is a component of the natural parameter.

Let us denote by $i_{\psi\psi}(\psi, \lambda)$ and $i_{\lambda\lambda}(\psi, \lambda)$ the components of the Fisher information matrices corresponding to ψ and λ respectively. Also, let $\hat{\psi}_i$ and $\hat{\lambda}_i$, $i = 1, 2$, be the maximum likelihood estimators based on the first and second set of samples, assumed to exist, and let $p_{\hat{\psi}_1, \hat{\lambda}_1}(\hat{\psi}_1, \hat{\lambda}_1)$ be the selection function in terms of $(\hat{\psi}_1, \hat{\lambda}_1)$. By orthogonality, the asymptotic distribution of $\hat{\psi}_1$ given $\hat{\lambda}_1$ and the selection event is selective normal with mean ψ , variance $\{n_1 i_{\psi\psi}(\psi, \lambda)\}^{-1}$, and selection function $p_{\hat{\psi}_1}(\hat{\psi}_1) = p_{\hat{\psi}_1, \hat{\lambda}_1}(\hat{\psi}_1, \hat{\lambda}_1)$, where $\hat{\lambda}_1$ is fixed at its observed value. Also, the asymptotic distribution of $\hat{\psi}_2$ given $\hat{\lambda}_2$ is $N(\psi, \{n_2 i_{\psi\psi}(\psi, \lambda)\}^{-1})$. Therefore, for a fixed value of λ , $(\hat{\psi}_1, \hat{\psi}_2) | (\hat{\lambda}_1, \hat{\lambda}_2)$ follows asymptotically a selective normal distribution, with variance-stabilising transformation $g(\psi; \lambda)$ satisfying $g'(\psi; \lambda) = i_{\psi\psi}(\psi, \lambda)^{1/2}$. Since $\hat{\lambda}_1$ is only mildly informative about ψ , the same heuristic argument as before would lead to adoption of the conditional prior $\pi(\psi | \lambda) \propto i_{\psi\psi}(\psi, \lambda)^{1/2} \pi_\nu\{g(\psi; \lambda)\}$, where $\pi_\nu(\nu)$ is a suitable non-informative prior for the selective normal-location model with location parameter $\nu = g(\psi; \lambda)$. Since this choice does not constrain the marginal prior of λ , a generic joint prior for (ψ, λ) would be of the form

$$\pi(\psi, \lambda) \propto c(\lambda) i_{\psi\psi}(\psi, \lambda)^{1/2} \pi_\nu\{g(\psi; \lambda)\},$$

where $c(\lambda)$ is an arbitrary prior density for λ .

4.3.1 Inference for a selected normal mean with an unknown variance

A simple selection problem involving nuisance parameters occurs when inference is required for a selected normal mean and the variance is unknown. Let $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown, but only μ is of direct interest to us. Suppose that the sample is divided into two sets of sizes n_1 and n_2 , with respective maximum likelihood estimators $\hat{\theta}_1 = (\bar{Y}_1, V_1)$ and $\hat{\theta}_2 = (\bar{Y}_2, V_2)$ of $\theta = (\mu, \sigma^2)$. Suppose that, in order to determine whether μ is of interest, we conduct the t -test $V_1^{-1/2}\bar{Y}_1 > n_1^{-1/2}t$ for some pre-specified value of t . The selective density of the data is

$$f_S(y_1, \dots, y_n; \mu, \sigma^2) = \frac{f(y_1, \dots, y_n; \mu, \sigma^2) \mathbf{1}(v_1^{-1/2}\bar{y}_1 > n_1^{-1/2}t)}{\mathbb{P}_{\mu, \sigma^2}(V_1^{-1/2}\bar{Y}_1 > n_1^{-1/2}t)}.$$

Note that in the selective model the marginal distribution of V_1 , with marginal density

$$f_S(v_1; \mu, \sigma^2) = \frac{f(v_1; \sigma^2) \mathbb{P}_{\mu, \sigma^2}(v_1^{-1/2}\bar{Y}_1 > n_1^{-1/2}t \mid v_1)}{\mathbb{P}_{\mu, \sigma^2}(V_1^{-1/2}\bar{Y}_1 > n_1^{-1/2}t)},$$

depends on μ , even though its non-selective density is free of it. Under selection, analogously to the random sample size example (Example 2.1.1). However, in this case the non-selective distribution of $\bar{Y}_1|V_1$ is not free of σ^2 , so there is no justification for conditioning on V_1 .

As expected by comparison with the univariate case, standard non-informative priors such as $\pi(\mu, \sigma^2) \propto \sigma^{-1}$ produce marginal posteriors for μ that overstate, on average, smaller values of μ . In this case, the proposed non-informative priors can be found easily, as the model is already parametrised orthogonally. For example, the option based on the Jeffreys prior takes the form

$$\pi_J(\mu, \sigma^2; v_1) \propto c(\sigma^2) \left\{ 1 + \frac{n_1}{n} h_2 \left(\frac{n_1^{1/2} \mu}{\sigma} - \frac{t v_1^{1/2}}{\sigma} \right) \right\}^{1/2},$$

where v_1 is the observed value of V_1 , and the option based on the probability-matching prior for $\gamma = 1$ is

$$\pi_y(\mu, \sigma^2; v_1) \propto c(\sigma^2) \left\{ 1 - \frac{h_1(\sigma^{-1} n_1^{1/2} \mu - \sigma^{-1} t v_1^{1/2})}{h_1(\sigma^{-1} n_1^{1/2} \mu - \sigma^{-1} n_1^{1/2} \bar{y})} \right\}.$$

The expression for $\gamma < 1$ is messy and therefore omitted, but can be obtained easily from (4.2). These priors assign lower probabilities to small values of μ relative to σ . Figure 4.5 illustrates the performance of $\pi_J(\mu, \sigma^2; v_1)$ for $c(\sigma^2) = \sigma^{-1}$ under repeated sampling from the selective model with $n_1 = 50$, $n_2 = 10$, $t = 2$, and true parameter

$(\mu_0, \sigma_0^2) = (0, 1)$. The prior $\pi_y(\mu, \sigma^2; v_1)$ gave numerical issues for small values of σ and was not considered. The plot shows the empirical distribution functions of $\Pi(\mu_0|Y_1, \dots, Y_n)$ for the prior $\pi(\mu, \sigma^2) \propto \sigma^{-1}$ and for the proposed prior, as estimated from 5×10^3 repetitions. The selection probability was computed by numerical evaluation of the integral

$$\mathbb{P}_{\mu, \sigma^2}(V_1^{-1/2}\bar{Y}_1 > n_1^{-1/2}t) = \int_0^\infty f(v_1; \sigma^2) \Phi\left\{\frac{n_1^{1/2}}{\sigma} \left(\mu - \frac{tv_1^{1/2}}{n_1^{1/2}}\right)\right\} dv_1,$$

and the marginal posterior distribution of μ was approximated with a Metropolis-Hastings algorithm with 5×10^3 steps. The results show that the selection-adjusted prior produces posterior inference with a more reliable frequentist calibration than the unadjusted one.

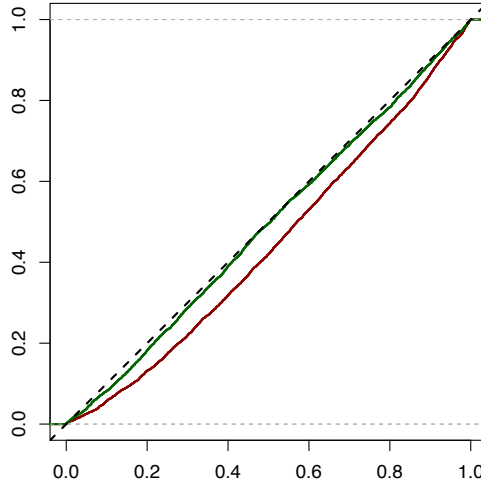


Figure 4.5: Estimated empirical CDFs of $\Pi(\mu_0|Y_1, \dots, Y_n)$ for the prior $\pi(\mu, \sigma^2) \propto \sigma^{-1}$ (red) and for Jeffreys prior (green).

4.3.2 Inference for the winner

Consider the standard problem of providing inference for the mean of a largest normal sample. Suppose we have n_1 observations from m different means, with sample means $Y_i \sim N(\theta_i, n_1^{-1})$, $i = 1, \dots, m$, and assume without loss of generality that $Y_1 = \max\{Y_i: i = 1, \dots, m\}$, so that θ_1 is the parameter of interest. Assume also that after selection we obtain another sample of size n_2 from θ_1 , with mean $\tilde{Y}_1 \sim N(\theta_1, n_2^{-1})$. We saw in Section 2.1 that there are two natural inferential models that are compatible with this selection rule. One has likelihood

$$L_1(\theta) \propto \frac{\phi\{n_2^{1/2}(\theta_1 - \tilde{y}_1)\} \prod_{i=1}^m \phi\{n_1^{1/2}(\theta_i - y_i)\}}{\mathbb{E}_\theta[\Phi\{n_1^{1/2}(\theta_1 - T)\}]}$$

where $n = n_1 + n_2$, $\theta = (\theta_1, \dots, \theta_m)^T$, and $T = \max\{Y_i: i = 2, \dots, m\}$, and another one has likelihood

$$L_2(\theta) \propto \frac{\phi\{n_2^{1/2}(\theta_1 - \tilde{y}_1)\} \prod_{i=1}^m \phi\{n_1^{1/2}(\theta_i - y_i)\}}{\Phi\{n_1^{1/2}(\theta_1 - t)\}}$$

which is derived from conditioning on the data from the non-selected means. For this model, the prior proposed in this section takes the form

$$\pi(\theta) \propto c(\theta_2, \dots, \theta_m) \left[1 + \frac{n_1}{n} h_2 \left\{ n_1^{1/2}(\theta_1 - t) \right\} \right]^{1/2} \quad (4.3)$$

in the case of the Jeffreys prior, and similarly for the probability-matching prior. Note that in this case the original parametrisation is also orthogonal.

Since the distribution of $Y_2, \dots, Y_m$ depends on θ_1 given selection, conditioning on them, even if advisable on conceptual grounds, will reduce the inferential power. In the following table we investigate this by comparing the length of credible intervals obtained by attaching prior (4.3) with $c(\theta_2, \dots, \theta_m) = 1$ to the likelihoods $L_1(\theta)$ and $L_2(\theta)$. We set $n_1 = n_2 = 5$, true parameter $\theta = (0, \dots, 0)^T$, and $m = 2, 5, 10, 20$. The credible intervals were computed by taking the lower and upper 0.05 quantiles of the marginal posterior distribution of θ_1 . To compute each of the intervals we run a Metropolis-Hastings algorithm with 10^4 steps. The simulation size was 5×10^3 , and the maximum standard error of the figures in the table was 0.005. We can see there is some loss of power when we condition on the data from the non-selected means, as indicated by longer intervals, and that this loss worsens slightly as the number of means increases. We note, however, that the simulation for $L_1(\theta)$ was much more computationally challenging than for $L_2(\theta)$ for large values of m . The results for $L_1(\theta)$ were found to be quite sensitive to the choice of integration boundaries in the approximation of the selection probability.

	$L_1(\theta)$		$L_2(\theta)$	
	Cov.	Length	Cov.	Length
$m = 2$	0.913	1.149	0.902	1.205
$m = 5$	0.899	1.193	0.902	1.292
$m = 10$	0.899	1.183	0.906	1.329
$m = 20$	0.887	1.148	0.903	1.352

Table 4.2: Coverage and average length of 0.9-credible intervals for the mean giving the largest sample, derived from likelihoods $L_1(\theta)$ and $L_2(\theta)$ and Jeffreys prior.

4.4 Appendix

4.4.1 Proof of Proposition 3

Without loss of generality we may assume that $\sigma^2 = 1$ and $t = 0$. Let $h_i(x)$ denote the i -th derivative of $\log \Phi(x)$ and, for $y > 0$, define

$$H(\theta; y) = \mathbb{P}_\theta(Y \geq y | Y > 0) = \frac{\Phi(\theta - y)}{\Phi(\theta)},$$

which is a distribution function as a function of θ , and

$$g(\theta; y) = -\frac{\frac{\partial}{\partial \theta} H(\theta; y)}{\frac{\partial}{\partial y} H(\theta; y)} = 1 - \frac{h_1(\theta)}{h_1(\theta - y)} > 0.$$

This function is strictly increasing. Indeed,

$$g'(\theta; y) = \frac{h_1(\theta)h_2(\theta - y)}{h_1(\theta - y)^2} - \frac{h_2(\theta)}{h_1(\theta - y)},$$

which is positive if

$$\frac{h_2(\theta - y)}{h_1(\theta - y)} > \frac{h_2(\theta)}{h_1(\theta)},$$

that is, if $h_2(x)/h_1(x)$ is strictly decreasing. Since $h_2(x) = -xh_1(x) - h_1(x)^2$, $h_2(x)/h_1(x) = -x - h_1(x)$, so $(\partial/\partial x)\{h_2(x)/h_1(x)\} = -1 - h_2(x) < 0$ for all $x \in \mathbb{R}$. The latter claim follows from the fact that $1 + h_2(x)$ is the variance of a $N(x, 1)$ distribution truncated to $(0, \infty)$. Since $g(\theta; y)$ is strictly increasing, if used as a formal prior density for θ , it satisfies

$$\Pi(\theta_0 | y) > \frac{\int_{-\infty}^{\theta_0} g(\theta; y) \Phi(\theta)^{-1} \phi(\theta - y) d\theta}{\int_{-\infty}^{\infty} g(\theta; y) \Phi(\theta)^{-1} \phi(\theta - y) d\theta}.$$

But, by definition, $g(\theta; y)$ is such that

$$g(\theta; y) \frac{\phi(\theta - y)}{\Phi(\theta)} = \frac{\partial}{\partial \theta} H(\theta; y),$$

so

$$\mathbb{P}_{\theta_0}\{\Pi(\theta_0 | Y) < \alpha | Y > 0\} > \mathbb{P}_{\theta_0}\{H(\theta_0, Y) < \alpha | Y > 0\} = \alpha,$$

since $H(\theta_0, Y)$ is uniformly distributed given selection.

4.4.2 Derivation of the probability-matching prior with data splitting

Recall from Section 3.1 that

$$p(y) = \Phi \left\{ \left(\frac{n}{\gamma^{-1} - 1} \right)^{1/2} (y - t) \right\}.$$

We also have that

$$\varphi(\theta) = \Phi\{n_1^{1/2}(\theta - t)\}.$$

The definition gives

$$\begin{aligned} \pi_y(\theta) \propto \frac{1}{\phi(n^{1/2}(y - \theta))} & \left\{ \int_y^\infty n(\tilde{y} - \theta) \phi(n^{1/2}(\tilde{y} - \theta)) p(\tilde{y}) d\tilde{y} \right. \\ & \left. - \frac{\varphi'(\theta)}{\varphi(\theta)} \int_y^\infty \phi(n^{1/2}(\tilde{y} - \theta)) p(\tilde{y}) d\tilde{y} \right\}. \end{aligned} \quad (4.4)$$

Using that

$$\int x \phi(x) \Phi(a + bx) dx = \frac{b}{d} \phi\left(\frac{a}{d}\right) \Phi\left(xt + \frac{ab}{d}\right) - \phi(x) \Phi(a + bx)$$

for any constants a and b (Owen, 1980), where $d = (1 + b^2)^{1/2}$, the first integral can be expressed as

$$\begin{aligned} \gamma^{1/2} \phi(n_1^{1/2}(\theta - t)) & \left\{ 1 - \Phi \left\{ \frac{n^{1/2}}{(1 - \gamma)^{1/2}} (y - \theta + \gamma(\theta - t)) \right\} \right\} \\ & + \phi(n^{1/2}(y - \theta)) \Phi \left\{ \frac{n_1^{1/2}}{(1 - \gamma)^{1/2}} (y - t) \right\}. \end{aligned}$$

Substituting this expression in (4.4) and reorganising the terms gives the claimed expression.

Chapter 5

Splitting methods for regression models

We consider the problem of providing valid inference in a sparse regression setting after variable selection. This problem has been studied from many angles. Here, we consider a simple alternative to data splitting based on randomising the response vector, which allows for higher selection and inferential power than the former and is applicable with an arbitrary selection rule. We provide a theoretical and empirical comparison of both methods and extend the randomisation approach to non-normal settings. Our investigations show that the gain in power can be substantial.

5.1 Introduction

Suppose we have data $Y \sim N(\mu, \sigma^2 I_n)$, where $\mu \in \mathbb{R}^n$, $\sigma^2 > 0$, and I_n is the $n \times n$ identity matrix. We assume that the components of μ are modelled as a function of p covariates, $\mu_i = g(x_{i1}, \dots, x_{ip})$ for some unknown g , and denote by $X = (x_{ij}) \in \mathbb{R}^{n \times p}$ a known, fixed design matrix. In many situations, it is suspected that only a few covariates are truly active, and a preliminary variable-selection step is performed to identify these. Having screened a set of potentially relevant variables, we may want to provide inference for the regression coefficients of the best linear approximation of μ in the selected model, or some other parameter depending on the output of the selection step.

If the same data that was used for selection is also used to provide inference for the selected parameter, standard inferential procedures are subject to selection bias and can provide unreliable results, typically leading to overestimation of effect sizes and underestimation

of the variance; see e.g. [Hong et al. \(2018\)](#). Data splitting techniques, whereby a portion of the data is reserved for uncertainty quantification, offer a simple yet effective way to circumvent this problem. Unfortunately, data splitting often leads to procedures with little power both for identifying the active covariates and for providing inference for the selected parameters. In this chapter, we study an alternative to data splitting, motivated by the work by [Tian and Taylor \(2018\)](#), which provides a more efficient way of splitting the sample information, resulting in more powerful procedures, and which is easy to apply with a general variable-selection method. Our objective is to provide a theoretical and empirical comparison between the two splitting strategies.

Data splitting is a popular tool in prediction problems, where the hold-out observations are used to assess the accuracy of a predictive model. When the goal is inference rather than prediction, a frequent criticism of data splitting is that different splits can produce different selected models, and therefore two people analysing the same data may end up answering different questions. While this is a valid concern, we stress that selection based on the full data is not free from some level of arbitrariness, as the selection process always involves subjective decisions, including the choice of the selection rule itself, and in many cases a random input independent of the data; see for example the methods of [Wasserman and Roeder \(2009\)](#); [Meinshausen and Bühlmann \(2010\)](#); and [Candès et al. \(2018\)](#). Nevertheless, it is important to keep the effect of the random components as low as possible, as failure to do so results in high uncertainty about the relevance of the selected variables. In some cases, the arbitrariness of data splitting can be removed by averaging the results over many splits, but combining the results from the different splits in an efficient way is challenging, as shown by [DiCiccio et al. \(2020\)](#).

The splitting technique considered in this chapter can be viewed as a variant of data splitting which produces datasets which are more similar to the full sample than those resulting from the latter method, and is therefore potentially less affected by randomness. Furthermore, it entails no extra computational cost with respect to data splitting. The method operates by applying the variable selection algorithm to a randomised version of the data, as in Section 1.3, and then basing inference on the conditional distribution of the data given its randomised form, thereby avoiding any selection bias. The general idea of basing selection on an artificial perturbation of the data in this context was proposed by [Tian and Taylor \(2018\)](#) as a way of deriving uniformly consistent and powerful inferential procedures, and has become a popular device in the literature. In the original paper, the authors compare the resulting inferential power of randomisation and data splitting in circumstances where it is possible to provide inference conditionally on the selection event, showing the superiority of the former. Here we compare the methods in circumstances where inference discards all the information of the selection split, and is therefore unaffected by the complexity of the selection rule. This is explained in more detail in Section

5.3, but the key point is that methods discussed here may be applied in circumstances where the conditional sampling distribution of the data given selection is intractable or unknown.

A large number of methods have been proposed in recent years to deal with selection bias in a regression setting. They can be broadly divided into two categories: those which assume that the selection algorithm is of a specific form; and those which provide guarantees for an arbitrary selection rule. An important class of methods in the first group are those based on the conditional approach, which are constructed by analysing the conditional distribution of the data given the output of the variable-selection algorithm, when this can be found explicitly. This line of work was started by [Lockhart et al. \(2014\)](#) and has subsequently been extended in multiple works such as [Lee and Taylor \(2014\)](#); [Loftus and Taylor \(2014\)](#); [Lee et al. \(2016\)](#); [Fithian et al. \(2017\)](#); [Tibshirani et al. \(2018\)](#); and [Panigrahi et al. \(2020\)](#). These methods often assume a normal generative model and an affine selection set (or a union of thereof), and follow from the realisation that, by conditioning on a carefully chosen set contained in the selection event, linear functions of the data follow a truncated normal distribution, for which exact inference is feasible. The second group of methods includes the POSI approach of [Berk et al. \(2013\)](#) and extensions of it ([Bachoc et al., 2017, 2020](#)), which achieve uniformly valid inference by maximising over all possible model selection procedures, and are very conservative as a result, as well as the data splitting approach of [Rinaldo et al. \(2019\)](#), which provides model-free procedures with asymptotically valid guarantees in a random-design setting. [Cox \(1975\)](#) analysed data splitting in a simple inferential problem involving many normal means and found it to be competitive against a natural alternative. Methods based on data splitting have also been considered in more recent works such as [Rubin et al. \(2006\)](#), [Wasserman and Roeder \(2009\)](#), [Ignatiadis et al. \(2016\)](#), and [DiCiccio et al. \(2020\)](#). [Fithian et al. \(2017\)](#) observed that inference after data splitting based only on the hold-out observations is inadmissible, being always dominated by *data-carving* rules, which consider the sampling distribution of the full data, conditional on the selection event. Such data-carving may, however, be very complicated or intractable to implement in many situations due to the complexity of the conditional distribution.

5.2 Inferential targets

Suppose that, for a given data vector $Y = y$, a variable-selection algorithm selects a subset $s \subseteq \{1, \dots, p\}$ of the covariates. In general, determining an appropriate inferential objective post-selection is not straightforward. Under a linearity assumption, $\mu = X\beta$ for some $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$, a natural possibility is to provide inference for the

components of β associated with the selected variables, $\{\beta_i: i \in s\}$. However, this is a difficult problem when $p > n$ as the model is not identifiable. A popular alternative target of inference is the projection parameter, proposed originally by [Berk et al. \(2013\)](#). The projection parameter is defined as the regression parameter of μ projected onto the subspace spanned by the selected columns of X :

$$\beta_s(X) = \arg \min_{z \in \mathbb{R}^{|s|}} \mathbb{E} [\|Y - X(s)z\|^2] = \{X(s)^T X(s)\}^{-1} X(s)^T \mu,$$

where $X(s)$ is the submatrix of X that contains the selected columns and $|s|$ denotes the number of selected covariates. In other words, it is the best linear predictor of μ in the selected model. When the model happens to be linear in the selected covariates, $\mu = X(s)\beta_s$ for some $\beta_s \in \mathbb{R}^{|s|}$, $\beta_s(X)$ is simply β_s . Otherwise, the interpretation of the projection parameter is less transparent: the j -th component of $\beta_s(X)$ may be viewed as the average change of the response when the j -th selected covariate increases by one unit, approximated in the selected model. An alternative interpretation is given in Section 5.6.4; see also Section 3.2 of [Berk et al. \(2013\)](#). When the full model is not linear, one can also consider the projection parameter using the full design matrix, $\beta^F(X) = (X^T X)^{-1} X^T \mu$, assuming X has full column rank. Some authors refer to this parameter as the full target, and to the previous one as the partial target. In this case, post-selection inference may be provided for the components of $\beta^F(X)$ associated with the selected variables, $\{\beta^F(X)_i: i \in s\}$. Another option, presumably more common in practice, is to proceed under the assumption that $\mu = X(s)\beta_s$ and carry out inference on β_s . In the random-design case, [Rinaldo et al. \(2019\)](#) develop inferential methods for other choices of the selected parameter which depend on the distribution of the covariates.

Let us denote a generic selected parameter, possibly depending on the design matrix as well as on the selected set, by $h_s(X)$. Since the design matrix is assumed fixed, the inferential problem can be cast as a selective inference one, with $\psi = h_s(X)$. Thus, by the conditional approach, inferential statements about $h_s(X)$ should be based on the conditional distribution of the data given $S = s$, where now the selection event S is identified by the random set of selected covariates. For simplicity, we will assume that the choice of the interest parameters depends only on s , so that inference on a given $h_s(X)$ is required if and only if $S = s$.

For some popular variable-selection algorithms, such as the lasso or stepwise procedures with fixed tuning parameters, the conditioning event $\{S = s\}$ can be studied analytically; see e.g. [Lee and Taylor \(2014\)](#); [Loftus and Taylor \(2014\)](#); and [Lee et al. \(2016\)](#). In most cases, however, this event is too complicated to be explored analytically. Furthermore, even when they can be implemented, conditional methods can be very conservative if we use all the data in the selection step ([Kivaranovic and Leeb, 2020](#)). In such cases, data

splitting offers an analytically simple and computationally light solution to the inference problem.

5.3 Splitting methods

The most common form of data splitting is simple data splitting. Here, a fraction $f = n_1/n$, $1 \leq n_1 < n$, is specified, and a set of indices r is chosen uniformly at random from the subsets of $\{1, \dots, n\}$ of size n_1 . Then, the sets of observations $\{Y_i: i \in r\}$ and $\{Y_i: i \in r^c\}$ are respectively used for selection and for inference, where r^c denotes the complement of r . Since all the observations are independent, the conditional distribution of the inference set given the output of the selection step is the same as the unconditional one, so classical procedures can be used to provide valid inference for a selected parameter. More elaborated splitting rules can be found in the prediction literature; see e.g. [Reit-ermanová \(2010\)](#). These rules allocate the samples to the selection and inferential sets according the observed values of the covariates, usually trying to divide them as evenly as possible, to ensure that the analyst has access to similar regions of the design space in both stages. For example, when the observations come from different groups, stratified data splitting rules allocate the samples so that both the inference and the selection sets contain observations from all the observed groups. Quite generally, then, a data splitting strategy can be formalised as a random variable R , possibly depending on X , taking values in the power set of $\{1, \dots, n\}$.

In the current context, data splitting techniques can be viewed as particular cases of a more flexible method. Suppose that W is a random quantity, possibly depending on X , and that in the selection step we only allow ourselves to observe the value of a function $U \equiv u(Y, W)$. Since selection depends on the data only through U , inference based on the conditional distribution of the data given its observed value, $Y \mid U$, is free of selection bias, and does not require knowledge of the selection mechanism (note that, by contrast, a conditional approach would base inference on $Y \mid \{S = s\}$). Abstractly, we may think of the data under this conditional distribution as the portion reserved for inference. Data splitting techniques can be written this way by letting $W = R$ and U be the set of observations allocated to selection. A natural question to examine, then, is whether there exist data decompositions of this form that are better than data splitting.

[Tian and Taylor \(2018\)](#) proposed randomisation schemes of the form $U = Y + \gamma W$, where W is n -dimensional artificial noise and $\gamma > 0$ is a tuning parameter which controls the amount of information reserved for inference. Small values of γ assign most of the sample information for selection, while large values allocate most of it for inference. One clear

advantage of this approach over data splitting is that it gives access to all the observed values of the covariates both at the selection and at the inferential stages, while in data splitting we only have access to a subset of them at each stage. To ensure high inferential power, [Tian and Taylor \(2018\)](#) recommend that the distribution of W has tails at least as heavy as the normal distribution. If the observation variance σ^2 is known, the choice $W \sim N(0_n, \sigma^2 I_n)$, where 0_n is an n -dimensional vector of zeroes, allows for a remarkably simple analysis. With this choice, $U \sim N(\mu, (1 + \gamma^2)\sigma^2 I_n)$ is of the same parametric form as the data. Furthermore, basing inference for the selected parameter on $Y | U$ amounts to basing it on the marginal distribution of $V = Y - \gamma^{-1}W \sim N(\mu, (1 + \gamma^{-2})\sigma^2 I_n)$. This follows because U and V are independent, as they are uncorrelated and normal, and jointly sufficient. Henceforth, we refer to this split as the (U, V) decomposition. [Tian \(2020\)](#) used such a decomposition to derive an unbiased estimator of the prediction risk in a related context.

If the error variance is unknown but can be estimated with reasonable precision, an approximate decomposition can be achieved by plugging in the variance estimate in the variance of W . In Section 5.5 we consider the asymptotic validity of this approach. In the linear case, if p is small relative to n , σ^2 can be estimated in the classical way. Otherwise we have to resort to high-dimensional alternatives. In our simulation studies we used the estimator implemented in the `selectiveInference` R package of [Tibshirani et al. \(2019\)](#), which estimates σ^2 using the residual sum of squares from a lasso fit, with the penalty parameter tuned by cross-validation. The good performance of this estimator in sparse models was demonstrated in [Reid et al. \(2016\)](#). Other methods for variance estimation in high dimensions include [Fan et al. \(2012\)](#) and [Bayati et al. \(2013\)](#).

5.4 Theoretical comparison

In the low-dimensional linear case, $\mu = X\beta$ and $p < n$, the optimality of a design X is commonly measured through summary statistics of the inverse Fisher information, $\sigma^2(X^T X)^{-1}$. Popular criteria in the literature include the D -optimality criterion, which minimises the determinant of the inverse information, the A -optimality criterion, which minimises the trace, and a few others. When a data split r is used to partition the sample, the inverse information matrices of the selection and inference sets are $\sigma^2(X_r^T X_r)^{-1}$ and $\sigma^2(X_{r^c}^T X_{r^c})^{-1}$, where X_r and X_{r^c} are the design matrices that contain only the observations in splits r and r^c , respectively. Correspondingly, an additive randomisation scheme $U = Y + \gamma W$, $W \sim N(0_n, \sigma^2 I_n)$, produces splits of the inverse Fisher information $(1 + \gamma^2)\sigma^2(X^T X)^{-1}$ and $(1 + \gamma^{-2})\sigma^2(X^T X)^{-1}$.

Proposition 4, which we state for a generic parametric model, shows that the latter split is more efficient than a class of data splitting strategies when the quality of the inverse Fisher information is measured in a particular way. For this analysis, we let $Y_1, \dots, Y_n$ be independent (not necessarily identically distributed) observations from a generic regular parametric model with p -dimensional parameter β , and write $I_i(\beta)$ for the Fisher information provided by the i -th observation, and $\mathcal{I}_Y(\beta) = \sum_{i=1}^n I_i(\beta)$ for the full-data Fisher information. A generic data split r distributes the information between the selection and inferential tasks as

$$\mathcal{I}_Y(\beta) = \sum_{i \in r} I_i(\beta) + \sum_{i \in r^c} I_i(\beta) \equiv \mathcal{I}_r(\beta) + \mathcal{I}_{r^c}(\beta).$$

The result concerns the behaviour of $\mathcal{I}_R(\beta)$ and $\mathcal{I}_{R^c}(\beta)$ under repeated sampling from a random data splitting strategy R , such as simple data splitting.

In what follows, all expectations are taken with respect to the distribution of R , and, for two positive definite matrices A and B , we use the notation $A < B$ ($\leq$) to indicate that $A - B$ is negative (semi)definite. Furthermore, we say that a real-valued matrix function κ is convex if it satisfies $\kappa\{tA + (1-t)B\} \leq t\kappa(A) + (1-t)\kappa(B)$ for all matrices A and B in its domain and $t \in (0, 1)$, and we say that it is strictly increasing if $A < B$ implies $\kappa(A) < \kappa(B)$. The proof of the proposition can be found in the appendix.

Proposition 4. Let R be any splitting strategy satisfying $\mathbb{P}(i \in R) = f$ for some $0 < f < 1$ and all $i = 1, \dots, n$, and let κ be a real-valued function defined on the set of $p \times p$ positive definite matrices which is convex and strictly increasing. Furthermore, assume that $\mathcal{I}_r(\beta)$ and $\mathcal{I}_{r^c}(\beta)$ are invertible for all r in the support of R , and that $\mathcal{I}_{r_1}(\beta) \neq \mathcal{I}_{r_2}(\beta)$ for some r_1 and r_2 . Then,

$$\kappa \left\{ \frac{1}{f} \mathcal{I}_Y(\beta)^{-1} \right\} < \mathbb{E} \left[\kappa \left\{ \mathcal{I}_R(\beta)^{-1} \right\} \right], \quad \kappa \left\{ \frac{1}{1-f} \mathcal{I}_Y(\beta)^{-1} \right\} < \mathbb{E} \left[\kappa \left\{ \mathcal{I}_{R^c}(\beta)^{-1} \right\} \right].$$

Examples of functions κ that satisfy the conditions of Proposition 4 include: $\kappa(A) = \text{tr}(A)$, related to the average variance of the estimates of the regression coefficients; $\kappa(A) = v^T A v$ for any vector $v \in \mathbb{R}^p$, related to the estimation variance of a specific combination of parameters $v^T \beta$; $\kappa(A) = \max\{\text{diag}(A)\}$, related to the maximum estimation variance of the regression coefficients; and $\kappa(A) = \lambda_{\max}(A)$, where $\lambda_{\max}(A)$ is the maximum eigenvalue of A , related to the variance estimation of the first principal component.

Proposition 4 says that, if we measure the quality of a design by such a κ , a splitting scheme $u(Y, W)$ which splits the Fisher information as $\mathcal{I}_Y(\beta) = f\mathcal{I}_Y(\beta) + (1-f)\mathcal{I}_Y(\beta)$, where $f\mathcal{I}_Y(\beta)$ and $(1-f)\mathcal{I}_Y(\beta)$ are the information matrices at the selection and inferential stages, is more efficient, on average, than any data splitting strategy for which every

sample has the same probability of being allocated to either set.

In additive models of the form $Y = \mu + \varepsilon$, splitting strategies satisfying the former condition are trivially given by rules of the form $U = Y + \gamma W$, where W has $\mathbb{E}[W] = 0_n$ and $\text{Var}(W) = \text{Var}(\varepsilon)$, and γ satisfies $(1 + \gamma^2)^{-1} = f$. On the other hand, the set of data splitting strategies for which $\mathbb{P}(i \in R)$ is constant include: simple data splitting with splitting fraction $f = n_1/n$; the class of stratified sampling rules where the observations are divided into k equal-size sets, and the selection set is formed by choosing a fixed number of observations uniformly at random from each set; and any splitting rule with $\mathbb{P}(R = A) = \mathbb{P}(R = A^c) = 1/2$ for some $A \subseteq \{1, \dots, n\}$. For example, we might use a rule that divides the samples into two sets as evenly as possible, and then label them as selection and inference sets at random.

5.5 Generalised (U, V) decomposition

In additive models, efficient splits are easily achievable via additive perturbation of the data. However, when the observation error and the randomisation noise are not normal, the distributions of U and $Y|U$ are generally not available in closed form, complicating the selection and inferential analyses. Furthermore, when the observation variance has to be estimated or the normality assumption is mildly violated, basing inference on the marginal distribution of V , as described above, is not formally justified. In this section we provide a set of conditions under which the (U, V) decomposition is asymptotically valid in more general settings.

Suppose that the model is $Y = \mu + \varepsilon$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, where the errors are a random sample from an unknown distribution, not necessarily normal. We assume that any selected parameter can be written in the form $\eta^T \mu$ for some contrast $\eta \in \mathbb{R}^n$ (note that all parameters considered in Section 5.2 can be written in this way). If the errors are $N(0, \sigma^2)$ and we know σ^2 , exact post-selection inference for $\eta^T \mu$ after selection based on $U = Y + \gamma W$ is provided via $\eta^T V \sim N(\eta^T \mu, (1 + \gamma^{-2})\sigma^2 \|\eta\|^2)$, where $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^n$.

Theorem 3 gives conditions under which this approach is asymptotically valid when the errors are not normal and the observation variance is unknown but can be estimated consistently. The result considers asymptotic regimes with a fixed number of covariates and a growing number of observations. Unless otherwise stated, all the elements involved in the analysis depend on the sample size n . This dependence is not made explicit to avoid saturating the notation. Also, for a vector or matrix A , $\max(A)$ denotes the maximum

absolute entry of A , and, for a square matrix A , $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A . The proof of the theorem can be found in the appendix.

Theorem 3. Let p be fixed, $n \rightarrow \infty$, and consider a sequence of selection sets $s_1, s_2, \dots \subseteq \{1, \dots, p\}$ such that the sequence selection probabilities, $\{\mathbb{P}(S = s_i) : i \in \mathbb{N}\}$, is bounded away from zero. Assume that all the selection events $\{S = s_i\}$ can be written in the form $\{X^T U \in E\}$, where $U = Y + \gamma W$ for a fixed $\gamma \in (0, 1)$, and E is a finite union of convex sets. Furthermore, assume that $X^T X$ is invertible for all n , $\lambda_{\max}\{(X^T X)^{-1}\} = O(n^{-1})$, $\max(X) = o(n^{1/6})$, and that the contrasts η satisfy $\max(\eta)\|\eta\|^{-1} = O(n^{-1/2})$. If the observation errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with mean zero, variance σ^2 , and $\mathbb{E}[|\varepsilon_1|^3] < \infty$, and $W = \hat{\sigma}Z$, where $\hat{\sigma}$ is a weakly consistent estimator of σ and $Z \sim N(0_n, I_n)$ is independent of the data, then

$$(1 + \gamma^{-2})^{-1/2} \sigma^{-1} \|\eta\|^{-1} (\eta^T V - \eta^T \mu) | \{S = s_n\} \xrightarrow{d} N(0, 1) \quad \text{as } n \rightarrow \infty,$$

where $V = Y - \gamma^{-1}W$.

The requirement that selection depends on the data only through $X^T U$ allows us to reduce the dimension of the conditioning event from n to p , which stays bounded. It is satisfied by all penalised-likelihood methods, which estimate the active set of covariates by the support of

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|U - X\beta\|^2 + l(\beta)$$

for an appropriate loss function l . This follows because the first term can be written, up to an additive constant independent of β , as $\|X(X^T X)^{-1} X^T U - X\beta\|^2$, which allows us to write the optimisation problem as

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|X(X^T X)^{-1} X^T U - X\beta\|^2 + l(\beta)$$

Regarding the convexity assumption, suppose that, for a given s , the selection event in terms of the full vector U is $E' \subseteq \mathbb{R}^n$. Under the previous condition, the selection event in terms of $X^T U$ is $E = X^T E' = \{X^T u : u \in E'\}$. Therefore, if $E' = E'_1 \cup \dots \cup E'_m$, with $E'_1, \dots, E'_m$ convex, $E = X^T E'_1 \cup \dots \cup X^T E'_m$ is also a union of convex sets, so the requirement can be imposed on E' directly. Verifying that E' is a union of convex sets can still be difficult problem in general, but it has been shown to be true for some widely used selection rules such as forward stepwise regression, the lasso, and least angle regression, all with fixed tuning parameters (Tibshirani et al., 2018). We note, however, that this requirement is mainly convenient from a mathematical perspective and can likely be weakened.

The asymptotic condition on η is very natural. It ensures that the asymptotic support of

$\eta/\|\eta\|$ is unbounded. When inference is sought for a selected coefficient or for a projection parameter, the contrasts are of the form $X(s)\{X(s)^T X(s)\}^{-1}e_j$ for some $s \subseteq \{1, \dots, p\}$ and $1 \leq j \leq p$, where $e_j \in \mathbb{R}^p$ is the j -th vector of the canonical basis. These vectors have

$$\begin{aligned}\|\eta\|^2 &= e_j^T \{X(s)^T X(s)\}^{-1} e_j \\ &\geq \lambda_{\min} [\{X(s)^T X(s)\}^{-1}] \\ &= \lambda_{\max} \{X(s)^T X(s)\}^{-1} \\ &\geq \lambda_{\max} \{X^T X\}^{-1},\end{aligned}$$

where we have used that $\lambda_{\max}\{X(s)^T X(s)\} \leq \lambda_{\max}\{X^T X\}$. Furthermore, by similar arguments and the Cauchy–Schwarz inequality, it follows that

$$\begin{aligned}\max(\eta) &= \max_{1 \leq i \leq n} e_i^T X(s) \{X(s)^T X(s)\}^{-1} e_j \\ &\leq \max_{1 \leq i \leq n} \|e_i^T X(s)\| \|\{X(s)^T X(s)\}^{-1} e_j\| \\ &\leq p^{1/2} \max(X) [e_j^T \{X(s)^T X(s)\}^{-2} e_j]^{1/2} \\ &\leq p^{1/2} \max(X) \lambda_{\max} [\{X(s)^T X(s)\}^{-1}] \\ &= \max(X) O(n^{-1})\end{aligned}$$

under the assumptions of the theorem. If we further assume that the entries of X are bounded as n grows, $\max(X) = O(1)$, putting all together it follows that the asymptotic condition on the contrast is satisfied whenever $\lambda_{\max}(X^T X) = O(n)$.

5.6 Simulation study

We compared data splitting with the (U, V) decomposition in the context of the normal linear model. The data generating process was $Y = X\beta + N(0_n, I_n)$, where $\beta \in \mathbb{R}^p$ is an unknown sparse vector of coefficients. At each replication of the simulations, the rows of the design matrix were generated as independent samples from the distribution $N(0_p, \Gamma)$, where $\Gamma \in \mathbb{R}^{p \times p}$ is a Toeplitz matrix with (i, j) entry $\rho^{|i-j|}$ for some $\rho \geq 0$. The observation variance, fixed at 1, was assumed unknown, and was estimated in the classical way when $p < n/4$, and using the high-dimensional alternative otherwise, as per Section 5.3. We compared the splitting strategies in terms of selection power, selection stability, and inferential power in low and high-dimensional settings. All the simulation results can be found in the appendix.

The data splitting strategy considered was the DUPLEX (Snee, 1977). In the even case ($f = 1/2$), the DUPLEX proceeds as follows. First, it finds the two covariate observations

that are farthest apart (in Euclidean distance) and assigns them to the selection set. Then, it finds the next two observations that are farthest apart and assigns them to the inference set. Finally, it allocates the remaining samples one at a time, rotating between the selection and inference sets, by selecting the remaining sample that is farthest apart from the rest of the observations in the given set, starting with the selection set. If the selection and inferential sets have unequal sizes ($f \neq 1/2$), the algorithm is applied until the smallest set is filled, and all the remaining observations are assigned to the other one. When the covariate observations are not uniformly distributed, this algorithm ensures that neither the selection nor the inferential splits are imbalanced. We note that this algorithm does not satisfy the condition of Proposition 4, but it is nevertheless dominated by the (U, V) decomposition, as we shall see.

Following the discussion of Section 5.4, for a given fraction f , we compared the performance of the data splitting algorithm with splitting fraction f with the performance of the randomised procedure with tuning parameter $\gamma = (f^{-1} - 1)^{1/2}$. Specifically, for a given f and a given dataset (Y, X) , we considered two procedures. The first bases selection on $(Y^{(1)}, X^{(1)})$ and inference on $(Y^{(2)}, X^{(2)})$, where $Y^{(1)} = (Y_i: i \in r)^T$, $X^{(1)} = (x_{ij}: i \in r, j = 1, \dots, p)$, $Y^{(2)} = (Y_i: i \in r^c)^T$, and $X^{(2)} = (x_{ij}: i \in r^c, j = 1, \dots, p)$. The second one bases selection on $(U, X) = (Y + \gamma W, X)$ and inference on $(V, X) = (Y - \gamma^{-1} W, X)$, with $W = \hat{\sigma} Z$, where $Z \sim N(0_n, I_n)$ is artificially generated noise independent of the data and $\hat{\sigma}^2$ is the estimate of model variance. We used the classical variance estimator for the cases with $p < n/4$ and the high-dimensional estimator of Tibshirani et al. (2019) otherwise.

Regarding selection, we considered two recently proposed algorithms: the fixed- X knockoff algorithm of Barber and Candès (2015) and the stability selector paired with the lasso, proposed by Meinshausen and Bühlmann (2010) and improved by Shah and Samworth (2013), as implemented in the R packages `knockoff` (Barber et al., 2020) and `stabs` (Hofner and Hothorn, 2017). These are selection rules for which conditional inference is analytically intractable and for which computational approaches are very demanding (Markovic et al., 2019), so they constitute an example where splitting approaches would likely be preferred in practice. These algorithms aim at identifying the set of active coefficients while keeping the number of false discoveries under control. The knockoff provides a guarantee on the false discovery rate, while stability selection controls the expected number of false discoveries. In all the simulations we set the false discovery rate of the knockoff algorithm at 0.3, and the expected number of false discoveries and cutoff threshold of the stability algorithm at 3 and 0.7 (see the package documentation for details). Since the fixed- X knockoff is only applicable in settings with fewer covariates than observations, it was only considered in the lower dimensional cases.

5.6.1 Selection power

In this simulation, we generated 10^3 triplets (β, Y, X) independently for each combination of the following parameters: $n = 200$; $\rho = 0, 0.5$; $f = 1/2, 3/4$ (correspondingly $\gamma = 1, 3^{-1/2}$); $p = 30, 50$ for the knockoff algorithm, and $p = 200, 1000$ for the stability selection algorithm. For each combination of the parameters and each repetition, the true β was generated by sampling 10 non-zero positions uniformly at random and filling them with independent random variables distributed uniformly in the set $\{-1, -0.9, -0.8, \dots, -0.1, 0.1, \dots, 0.9, 1\}$. Then, for each β , a pair (Y, X) was generated, and the corresponding selection algorithm was applied to $(Y^{(1)}, X^{(1)})$ in the case of data splitting, and to (U, X) in the case of the randomised procedure.

The selection ability of the splitting methods was compared according to two criteria: true positive rate and power. The true positive rate is the average number of correct discoveries divided by the total number of active covariates, and the power is defined, for each possible value of $|\beta_i| = 0.1, \dots, 1$, as the average number of times a coefficient with absolute value $|\beta_i|$ is selected, averaged over all the generated coefficients of all β 's. We also computed, for comparison, the results corresponding to applying the selection algorithms to the full dataset, (Y, X) .

Table 5.1 shows the observed true positive rates of data-splitting and randomisation divided by the observed true positive rates of full-data selection, and Figures 5.1 and 5.2 show the empirical power functions of the three methods. The results clearly favour the randomised split procedure over data splitting, particularly in the cases with larger values of p . Quite remarkably, despite the fact that the choices of f and γ were balanced (in the sense described in Section 5.3), selection based on a randomised split yielded a performance which was in some cases closer to full-data selection than to data-splitting selection.

5.6.2 Selection stability

In addition to enjoying high power, it is important that the selection method does not depend strongly on ancillary components of the analysis: in the case of data splitting, on the choice of the selection and inference sets, and, in the case of randomisation, on the observed value of the artificial noise W . To measure the stability of the selection strategies with respect to these elements, a simulation was conducted in which the selection algorithms were applied to multiple splits of the same dataset. Since the DUPLEX is a deterministic allocation rule, in this section we considered a simple data splitting scheme instead.

We set $\beta = (1, 0.9, \dots, 0.1, 0, \dots, 0)^T$, $\rho = 0.5$, $f = 1/2, 3/4$, $p = 50$ for the knockoff, and $p = 400$ for stability selection. For each (f, p) , we generated 100 pairs (Y, X) and, for each pair, we sampled 50 selection sets uniformly at random and 50 realisations of the randomisation noise W , and applied the selection algorithms to each data split and perturbed instance of the data. For each (Y, X) , we recorded the average number of times each active covariate was selected across the different splits, an estimate of $\mathbb{P}(i \in S|Y, X)$, $i = 1, \dots, 10$. The empirical mean and standard deviations of the 100 estimated averages can be found in Table 5.2.

Randomisation was more stable than data splitting, because the corresponding selection probabilities of the active covariates were more concentrated around higher values. For example, for the knockoff with $f = 1/2$, for most values of (Y, X) we had more than a 90% probability of selecting β_3 under repeated sampling from W , as opposed to the $\sim 77\%$ of data splitting.

5.6.3 Inference for selected coefficients

In this simulation, we consider the problem of constructing confidence intervals for the selected coefficients $\{\beta_i : i \in s\}$. Firstly, we show that a face-value approach, which reports the standard confidence intervals ignoring selection, undercovers coefficients with small effect size $|\beta_i|$, while the splitting techniques provide valid intervals. This gives an illustration of the need to take selection into account in the inferential stage. Secondly, we compare the lengths of the intervals derived from the randomised procedure with respect to those obtained by data splitting. To avoid complications related to confidence interval construction in non-identifiable settings, we shall only consider here low-dimensional cases.

Assume that X has full rank and let $\hat{\beta}_i = e_i^T (X^T X)^{-1} X^T Y$ be the ordinary least squares estimator of β_i . In a classical, non-selective setting, where the inferential goals are determined prior to the data collection, marginal inference for β_i is based on the pivot $T_i = (\hat{\beta}_i - \beta_i)/\hat{\sigma}$, which follows a scaled t distribution with $n - p$ degrees of freedom, where $\hat{\sigma}^2$ is the classical variance estimator. If, however, inference on a given β_i is only provided for some data samples, T_i is no longer pivotal, and the resulting confidence intervals can be miscalibrated. To exemplify this, we run the following simulation. We fixed $n = 200$, $p = 30$, true parameter $\beta = (1, -1, 0.5, -0.5, 0.2, -0.2, 0, 0, \dots, 0)^T$, $\rho = 0, 0.5$, and $f = 1/2, 3/4$, and for each combination of (ρ, f) , we generated 5×10^3 pairs (Y, X) . We then applied the selection algorithm to each dataset as in the previous sections and constructed classical equal-tailed confidence intervals based on T_i for each selected coefficient, with a nominal coverage of 0.9.

Tables 5.3 and 5.4 contain the observed coverages of these intervals in the rows indicated by FV (face-value), averaged across coefficients with equal absolute effect. Clearly, the face-value intervals are unreliable when $|\beta_i| = 0$ or 0.2, with actual coverages significantly lower than the nominal one in most cases. Note that here the face-value approach is being implemented after selection based on a fraction of the sample information, so the selection effect is not as large as if all the data had been used for selection.

The rows labelled with HD contain the coverages of the intervals constructed using only the hold-out split of the data. In the case of data splitting, the intervals were derived from the t -pivots of the hold-out observations, $\hat{\beta}_i^{\text{DS}} = e_i^T (X^{(2)T} X^{(2)})^{-1} X^{(2)T} Y^{(2)}$. With the current simulation parameters, the number of remaining observations is always larger than the number of covariates, so $X^{(2)}$ has full rank almost surely. For the randomised procedure, we used the approximation $\hat{\beta}_i^{\text{R}} = e_i^T (X^T X)^{-1} X^T V \sim N(\beta_i, (1 + \gamma^{-2}) \sigma^2 e_i^T (X^T X)^{-1} e_i)$. The intervals were thus constructed by studentisation of $\hat{\beta}_i^{\text{R}}$. Despite the distributional approximation, we see that the coverages of the resulting confidence intervals were very close to the nominal one.

For the hold-out methods we also recorded the average length of the intervals, which can be found in Tables 5.5 and 5.6. The intervals provided by the randomised procedure were always shorter, on average, than those provided by the data splitting procedure. In the cases where the amount of information reserved for inference was small ($f = 3/4$), the matrices $X^{(2)}$ were of dimension 50×30 , and the resulting intervals were very wide as a consequence. The randomised procedure performed significantly better in these cases, giving intervals roughly 75% shorter than data splitting. The maximum observed standard deviations of the figures shown in the tables were, respectively, 0.009, 0.019, 0.004, and 0.007.

5.6.4 Inference for projection parameters

We now consider settings with a number of covariates exceeding the sample size. In these situations the full model is not identifiable, and the methods used in the previous section are not applicable. Instead, we consider the problem of constructing confidence intervals for the coefficients of a projection parameter. As noted by [Berk et al. \(2013\)](#), the coefficients of a projection parameter for a selected set s can be interpreted as a measure of the correlation between the selected coefficients and the mean μ , accounting for all the other selected covariates. This follows because $\beta_s(X)_i$ can be written as the projection of μ onto the residual vector of the i -th selected covariate regressed against the other selected covariates. Indeed, if $X(s)_i$ is the i -th column of $X(s)$ and $X(s)_{-i}$ is $X(s)$ with the i -th

column removed, a simple calculation shows that

$$\beta_s(X)_i = \frac{1}{\|r_i\|^2} r_i^T \mu, \quad r_i = \left\{ I_n - X(s)_{-i} (X(s)_{-i}^T X(s)_{-i})^{-1} X(s)_{-i}^T \right\} X(s)_i.$$

Thus, $\beta_s(X)_i$ provides a measure of the relevance of the i -th covariate relative to the selected set, with a null value indicating that the variable is redundant and can be safely removed from s .

If s had been fixed in advance and we knew the value of σ^2 , inference on the coefficients of $\beta_s(X)$ would be based on the components of $\hat{\beta}_s(X) = \{X(s)^T X(s)\}^{-1} X(s)^T Y \sim N(\beta_s(X), \sigma^2 \{X(s)^T X(s)\}^{-1})$. So, for a nominal coverage of $1 - \alpha$, the confidence intervals would be given by $[\hat{\beta}_s(X)_i \mp q_{1-\alpha/2} \sigma [e_i^T \{X(s)^T X(s)\}^{-1} e_i]^{1/2}]$, where $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. When σ^2 is unknown, we can plug-in the high-dimensional estimate used for the (U, V) decomposition, $\hat{\sigma}_{\text{HD}}^2$, say, and report $[\hat{\beta}_s(X)_i \mp q_{1-\alpha/2} \hat{\sigma}_{\text{HD}} [e_i^T \{X(s)^T X(s)\}^{-1} e_i]^{1/2}]$.

When s is data-dependent, hold-out inference can be provided similarly, with the difference that, for data splitting, the inferential target is not the full-projection parameter $\beta_s(X)$, but the projection parameter based on the hold-out observations. Inference is thus provided for $\beta_s(X^{(2)}) = \{X^{(2)}(s)^T X^{(2)}(s)\}^{-1} X^{(2)}(s)^T \mu_2$, where $\mu_2 = \mathbb{E}[Y^{(2)}]$, and is based on $\hat{\beta}_s^{\text{DS}}(X^{(2)}) = \{X^{(2)}(s)^T X^{(2)}(s)\}^{-1} X^{(2)}(s)^T Y^{(2)} \sim N(\beta_s(X^{(2)}), \sigma^2 \{X^{(2)}(s)^T X^{(2)}(s)\}^{-1})$. In the case of randomisation, the target parameter is still $\beta_s(X)$, and, as before, inference is based on the normal approximation to the distribution of

$$\hat{\beta}_s^{\text{R}}(X) = \{X(s)^T X(s)\}^{-1} X(s)^T V \sim N(\beta_s(X), (1 + \gamma^{-2}) \sigma^2 \{X(s)^T X(s)\}^{-1}).$$

In both cases, σ^2 was approximated by $\hat{\sigma}_{\text{HD}}^2$.

The simulation parameters were set as in Section 5.6.3 except for the number of covariates, which was $p = 400$. Here the knockoff was not considered, as it is not applicable with $p > n$. The results can be found in Tables 5.7 and 5.8. Note that in this case the columns indicate the absolute value of the full-model coefficients β_i , not the coefficients of the projection parameters, which in general depend on s . The coverage results were similar to the lower dimensional case, with the difference that the effect of selection was more pronounced here. In one of the cases considered, the coefficients of the projection parameters associated with the null covariates in the full model were almost guaranteed to be missed by the face-value intervals. The hold-out intervals, on the other hand, remained well-calibrated across the different coefficients. Regarding interval lengths, in this case we computed the average of all the intervals produced for a given selection set s , and averaged the results over selection sets of equal size in order to establish a more equitable comparison. We see that the average lengths were very similar in this case.

Recall, however, that the selection power of the latter method is substantially higher in high-dimensional settings, so in conjunction randomisation dominates data splitting. The maximum standard deviation of the coverage figures were 0.014, 0.045, 0.012 and 0.004, for $|\beta_i| = 0, 0.2, 0.5$, and 1, respectively. The maximum standard deviation of the length figures was 0.04. The entries with a dash indicate that no selection set of the corresponding size was selected in the simulation.

5.7 Appendix

5.7.1 Simulation results

Knockoff			Split	
f	ρ	p	DS	R
1/2	0	30	0.903	0.942
1/2	0	50	0.738	0.923
1/2	0.5	30	0.820	0.914
1/2	0.5	50	0.648	0.890
3/4	0	30	0.972	0.978
3/4	0	50	0.935	0.971
3/4	0.5	30	0.940	0.964
3/4	0.5	50	0.890	0.970

Stability			Split	
f	ρ	p	DS	R
1/2	0	200	0.694	0.867
1/2	0	1000	0.486	0.821
1/2	0.5	200	0.691	0.858
1/2	0.5	1000	0.478	0.820
3/4	0	200	0.895	0.948
3/4	0	1000	0.826	0.933
3/4	0.5	200	0.886	0.948
3/4	0.5	1000	0.826	0.934

Table 5.1: True positive rate of the selection algorithms applied after data splitting (DS) and randomisation (R), normalised by the true positive rate of selection applied to the full dataset.

		β_i									
f	Split	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Knockoff											
$1/2$	DS	0.81 (0.08)	0.76 (0.09)	0.77 (0.09)	0.74 (0.09)	0.71 (0.09)	0.67 (0.11)	0.62 (0.11)	0.53 (0.16)	0.38 (0.16)	0.15 (0.13)
	R	0.98 (0.06)	0.92 (0.18)	0.98 (0.06)	0.95 (0.10)	0.95 (0.11)	0.92 (0.13)	0.89 (0.13)	0.77 (0.23)	0.57 (0.26)	0.25 (0.22)
$3/4$	DS	0.98 (0.14)	0.95 (0.22)	0.97 (0.17)	0.97 (0.17)	0.96 (0.20)	0.88 (0.33)	0.89 (0.31)	0.80 (0.40)	0.62 (0.49)	0.25 (0.44)
	R	0.99 (0.04)	0.93 (0.19)	0.99 (0.04)	0.97 (0.10)	0.98 (0.08)	0.96 (0.10)	0.96 (0.11)	0.86 (0.22)	0.69 (0.29)	0.34 (0.32)
Stability											
$1/2$	DS	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	0.99 (0.02)	0.95 (0.10)	0.87 (0.17)	0.58 (0.28)	0.15 (0.18)	0.01 (0.02)
	R	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	0.98 (0.05)	0.95 (0.10)	0.76 (0.25)	0.30 (0.27)	0.04 (0.07)
$3/4$	DS	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.14)	0.88 (0.33)	0.31 (0.47)	0.08 (0.27)
	R	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	1.00 (0.02)	0.90 (0.20)	0.44 (0.37)	0.06 (0.14)

Table 5.2: Means and standard deviations of the estimated selection probabilities for fixed values of (Y, X) .

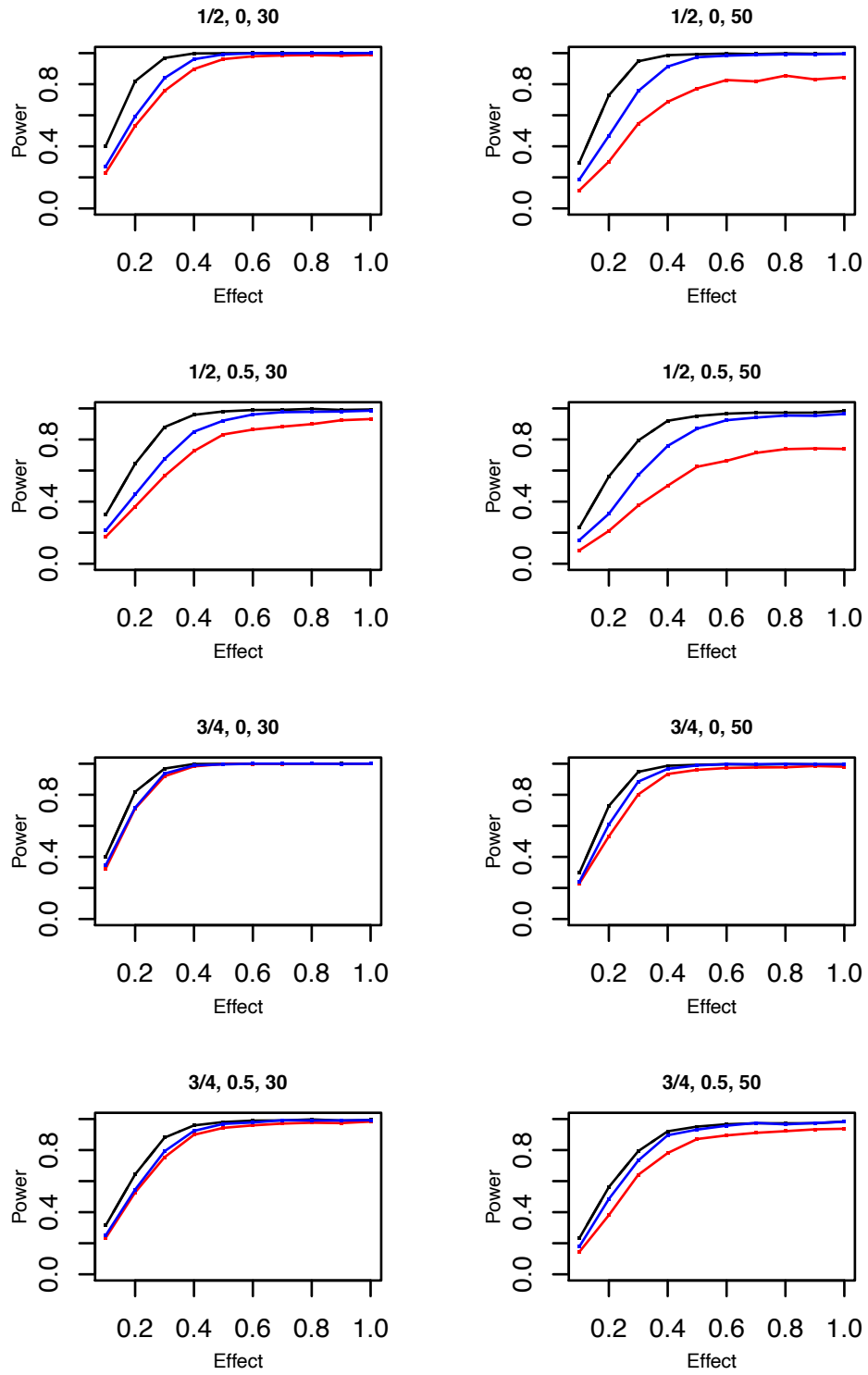


Figure 5.1: Power of the knockoff applied to the complete dataset (black), a data split (red), and a randomised version of the data (blue). Titles indicate (f, ρ, p) .

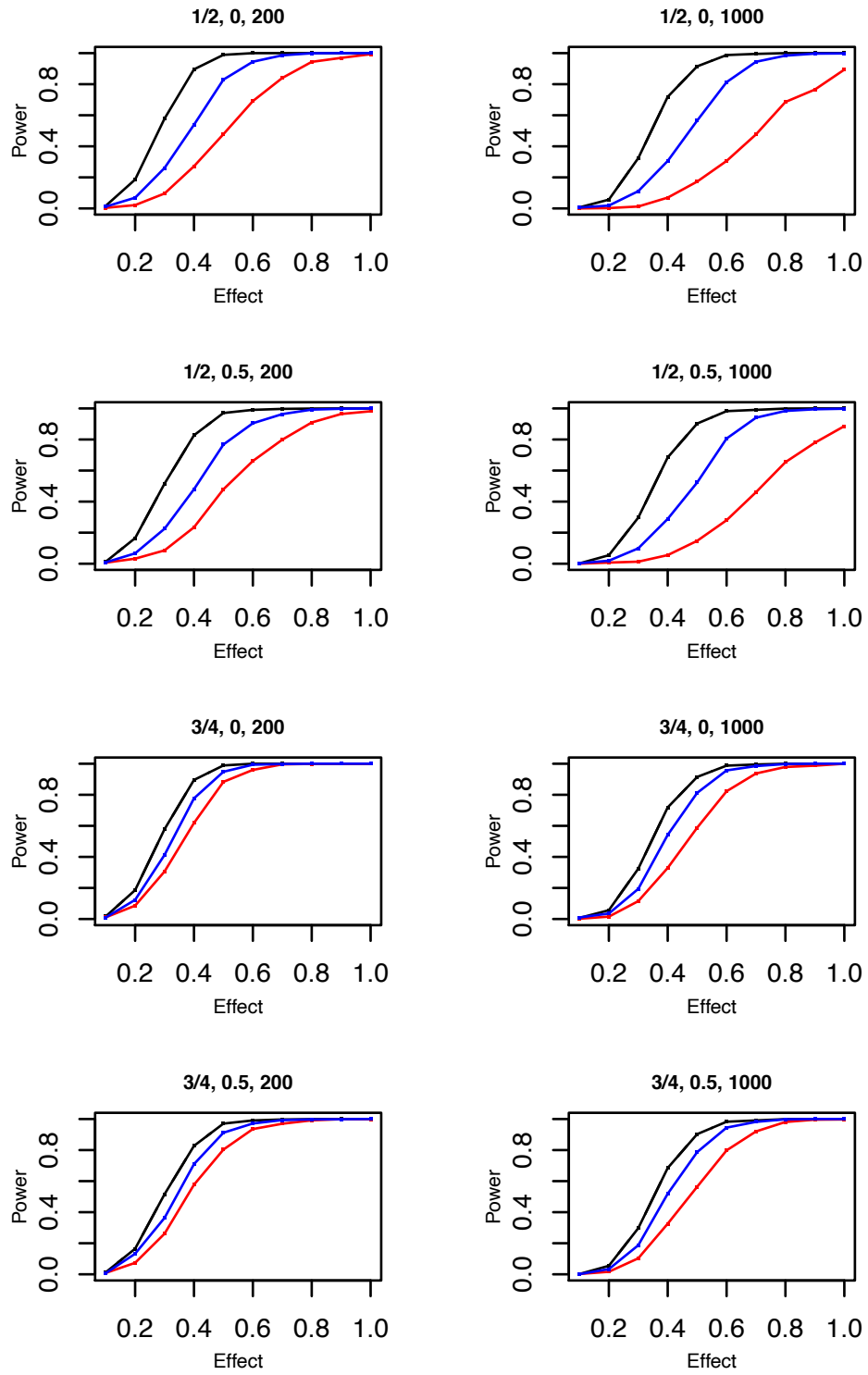


Figure 5.2: Power of stability selection applied to the complete dataset (black), a data split (red), and a randomised version of the data (blue). Titles indicate (f, ρ, p) .

f	ρ	Split	Method	$ \beta_i $			
				0	0.2	0.5	1
1/2	0	DS	FV	0.687	0.901	0.908	0.898
		R	FV	0.673	0.899	0.906	0.898
		DS	HD	0.898	0.913	0.900	0.900
		R	HD	0.896	0.899	0.899	0.903
1/2	0.5	DS	FV	0.739	0.852	0.910	0.899
		R	FV	0.720	0.840	0.912	0.897
		DS	HD	0.900	0.908	0.899	0.902
		R	HD	0.898	0.892	0.900	0.904
3/4	0	DS	FV	0.592	0.913	0.905	0.898
		R	FV	0.574	0.918	0.904	0.898
		DS	HD	0.903	0.897	0.897	0.896
		R	HD	0.898	0.902	0.900	0.896
3/4	0.5	DS	FV	0.664	0.858	0.915	0.900
		R	FV	0.650	0.852	0.914	0.899
		DS	HD	0.906	0.901	0.904	0.898
		R	HD	0.892	0.904	0.897	0.896

Table 5.3: Coverages of confidence intervals for the selected coefficients; knockoff selection.

f	ρ	Split	Method	$ \beta_i $			
				0	0.2	0.5	1
1/2	0	DS	FV	0.391	0.802	0.911	0.898
		R	FV	0.349	0.826	0.911	0.898
		DS	HD	0.903	0.906	0.900	0.900
		R	HD	0.901	0.894	0.898	0.897
1/2	0.5	DS	FV	0.617	0.816	0.895	0.900
		R	FV	0.558	0.801	0.895	0.899
		DS	HD	0.906	0.925	0.899	0.901
		R	HD	0.908	0.896	0.905	0.899
3/4	0	DS	FV	0.219	0.837	0.907	0.898
		R	FV	0.142	0.850	0.906	0.898
		DS	HD	0.908	0.902	0.896	0.896
		R	HD	0.885	0.904	0.899	0.903
3/4	0.5	DS	FV	0.469	0.799	0.903	0.899
		R	FV	0.400	0.808	0.906	0.899
		DS	HD	0.889	0.897	0.903	0.898
		R	HD	0.908	0.910	0.896	0.903

Table 5.4: Coverages of confidence intervals for the selected coefficients; stability selection.

				$ \beta_i $			
f	ρ	Split	Method	0	0.2	0.5	1
1/2	0	DS	HD	0.391	0.390	0.391	0.391
		R	HD	0.356	0.355	0.358	0.358
1/2	0.5	DS	HD	0.508	0.507	0.510	0.482
		R	HD	0.457	0.455	0.459	0.436
3/4	0	DS	HD	0.653	0.651	0.650	0.651
		R	HD	0.503	0.502	0.506	0.506
3/4	0.5	DS	HD	0.899	0.904	0.903	0.846
		R	HD	0.644	0.645	0.650	0.617

Table 5.5: Average length of confidence intervals for the selected coefficients; knockoff selection.

				$ \beta_i $			
f	ρ	Split	Method	0	0.2	0.5	1
1/2	0	DS	HD	0.389	0.391	0.391	0.392
		R	HD	0.360	0.354	0.358	0.358
1/2	0.5	DS	HD	0.506	0.506	0.509	0.483
		R	HD	0.459	0.458	0.457	0.438
3/4	0	DS	HD	0.657	0.648	0.651	0.651
		R	HD	0.507	0.500	0.507	0.507
3/4	0.5	DS	HD	0.890	0.904	0.902	0.847
		R	HD	0.648	0.648	0.646	0.620

Table 5.6: Average length of confidence intervals for the selected coefficients; stability selection.

f	ρ	Split	Method	$ \beta_i $			
				0	0.2	0.5	1
1/2	0	DS	FV	0.386	0.709	0.912	0.894
		R	FV	0.250	0.716	0.911	0.892
		DS	HD	0.897	0.861	0.894	0.900
		R	HD	0.879	0.895	0.898	0.899
1/2	0.5	DS	FV	0.417	0.750	0.839	0.906
		R	FV	0.325	0.753	0.875	0.908
		DS	HD	0.912	0.880	0.880	0.904
		R	HD	0.884	0.889	0.892	0.902
3/4	0	DS	FV	0.117	0.712	0.909	0.891
		R	FV	0.055	0.728	0.903	0.891
		DS	HD	0.882	0.898	0.900	0.900
		R	HD	0.894	0.903	0.900	0.899
3/4	0.5	DS	FV	0.183	0.713	0.885	0.905
		R	FV	0.129	0.701	0.899	0.901
		DS	HD	0.900	0.923	0.914	0.905
		R	HD	0.890	0.913	0.904	0.903

Table 5.7: Coverages of confidence intervals for the coefficients of the projection parameters; stability selection.

f	ρ	Split	Method	$ s $							
				1	2	3	4	5	6	7	8
1/2	0	DS	HD	-	0.34	0.34	0.34	0.33	0.33	0.34	-
		R	HD	-	0.35	0.34	0.33	0.33	0.33	0.33	0.30
1/2	0.5	DS	HD	0.34	0.38	0.37	0.37	0.36	-	-	-
		R	HD	0.35	0.39	0.37	0.36	0.36	0.35	0.34	-
3/4	0	DS	HD	-	0.48	0.48	0.48	0.48	0.47	0.47	0.48
		R	HD	-	0.52	0.49	0.47	0.47	0.46	0.46	0.45
3/4	0.5	DS	HD	0.48	0.55	0.54	0.52	0.52	0.53	-	-
		R	HD	0.49	0.56	0.54	0.52	0.51	0.48	0.49	-

Table 5.8: Average length of confidence intervals for the coefficients of the projection parameters; stability selection.

5.7.2 Proof of Proposition 4

Let $\mathcal{R}$ be the support of R , write $p_r = \mathbb{P}(R = r)$ for $r \in \mathcal{R}$, and let $1(\cdot)$ denote the indicator function. By the harmonic-arithmetic inequality for Hermitian matrices ([Bhagwat and Subramanian, 1978](#)), it follows that

$$\begin{aligned}
\frac{1}{f} \mathcal{I}_Y(\beta)^{-1} &= \left\{ \sum_{i=1}^n \mathbb{P}(i \in R) I_i(\beta) \right\}^{-1} \\
&= \left\{ \sum_{i=1}^n \sum_{r \in \mathcal{R}} p_r 1(i \in r) I_i(\beta) \right\}^{-1} \\
&= \left\{ \sum_{r \in \mathcal{R}} p_r \sum_{i=1}^n 1(i \in r) I_i(\beta) \right\}^{-1} \\
&= \left\{ \sum_{r \in \mathcal{R}} p_r \mathcal{I}_r(\beta) \right\}^{-1} \\
&< \sum_{r \in \mathcal{R}} p_r \mathcal{I}_r(\beta)^{-1} \\
&= \mathbb{E} \{ \mathcal{I}_R(\beta)^{-1} \}.
\end{aligned}$$

The inequality is strict because the $\mathcal{I}_r(\beta)$'s are not all equal. By the monotonicity of κ , this implies that

$$\kappa \left\{ \frac{1}{f} \mathcal{I}_Y(\beta)^{-1} \right\} < \kappa \left[\mathbb{E} \{ \mathcal{I}_R(\beta)^{-1} \} \right].$$

Furthermore, since κ is convex, Jensen's inequality gives

$$\kappa \left[\mathbb{E} \{ \mathcal{I}_R(\beta)^{-1} \} \right] \leq \mathbb{E} \left[\kappa \{ \mathcal{I}_R(\beta)^{-1} \} \right],$$

which shows the first inequality. The second one can be shown with a analogous argument.

5.7.3 Proof of Theorem 3

Assume without loss of generality that the maximum absolute entry of η is 1 for all n . Otherwise we can always divide it by its maximum absolute entry. Furthermore, assume for the moment that $\hat{\sigma} = \sigma$ and that E is convex for all n .

In the first part of the proof we show that, under these conditions, the standardised joint distribution of $X^T U$ and $\eta^T V$ is asymptotically normal. Define the $2n$ vector and $2n \times (p+1)$ matrix

$$T = \begin{pmatrix} \varepsilon \\ W \end{pmatrix}, \quad A = \begin{pmatrix} X & \eta \\ \gamma X & -\gamma^{-1} \eta \end{pmatrix}.$$

Note that A has full column rank and that

$$A^T A = \begin{pmatrix} (1 + \gamma^2)X^T X & 0_p \\ 0_p^T & (1 + \gamma^{-2})\|\eta\|^2 \end{pmatrix}, \quad A^T T = \begin{pmatrix} X^T U \\ \eta^T V \end{pmatrix} - \begin{pmatrix} X^T \mu \\ \eta^T \mu \end{pmatrix}.$$

Define also the $(p + 1)$ -dimensional vectors $P_i = \sigma^{-1}(A^T A)^{-1/2} A^T e_i \varepsilon_i$ for $i = 1, \dots, n$, and $P_i = \sigma^{-1}(A^T A)^{-1/2} A^T e_i W_i$ for $i = n + 1, \dots, 2n$, where e_i is the i -th canonical vector of $\mathbb{R}^{2n}$. Clearly, $\mathbb{E}[P_i] = 0_{p+1}$ for all $i = 1, \dots, 2n$. Furthermore,

$$\sum_{i=1}^{2n} P_i = \sigma^{-1}(A^T A)^{-1/2} A^T T \equiv Q, \quad \sum_{i=1}^{2n} \text{Var}(P_i) = I_{p+1}.$$

Consider events of the form $\{X^T U \in E, \eta^T V \leq c\}$ for $c \in \mathbb{R}$. These events can be equivalently written as $\{Q \in B_c\}$, where

$$B_c = \left\{ x \in \mathbb{R}^{p+1} : \sigma(A^T A)^{1/2} x + \begin{pmatrix} X^T \mu \\ \eta^T \mu \end{pmatrix} \in E \times (-\infty, c] \right\}.$$

Clearly, since E is convex, B_c is also convex. Thus, by a version of the multivariate Berry–Esseen theorem for convex sets (Raić, 2019, Theorem 1.1), it follows that

$$\begin{aligned} & |\mathbb{P}\{Q \in B_c\} - \mathbb{P}\{N(0_{p+1}, I_{p+1}) \in B_c\}| \\ & \leq \frac{\alpha}{\sigma^3} \left\{ \mathbb{E}[|\varepsilon_1|^3] \sum_{i=1}^n (e_i^T A(A^T A)^{-1} A^T e_i)^{3/2} + \mathbb{E}[|W_1|^3] \sum_{i=n+1}^{2n} e_i^T A(A^T A)^{-1} A^T e_i)^{3/2} \right\} \\ & = \tilde{\alpha} \sum_{i=1}^{2n} (e_i^T A(A^T A)^{-1} A^T e_i)^{3/2}, \end{aligned}$$

for some constants α and $\tilde{\alpha}$ independent of n . By the min-max theorem, we can bound the terms of the last sum as follows:

$$e_i^T A(A^T A)^{-1} A^T e_i \leq \|e_i^T A\|^2 \lambda_{\max} \left\{ (A^T A)^{-1} \right\} \leq (p + 1) \max(A)^2 \lambda_{\max} \left\{ (A^T A)^{-1} \right\}.$$

Hence,

$$\sum_{i=1}^{2n} (e_i^T A(A^T A)^{-1} A^T e_i)^{3/2} \leq 2n(p + 1)^{3/2} \max(A)^3 \lambda_{\max} \left\{ (A^T A)^{-1} \right\}^{3/2} = o(1)$$

by the various asymptotic conditions on X and η . Recall that we are assuming that the entries of η are bounded, and note that the eigenvalues of $A^T A$ are those of $X^T X$ multiplied by $1 + \gamma^2$, together with $(1 + \gamma^{-2})\|\eta\|^2$. The bound given by the theorem is uniform over convex sets, so we have shown that

$$|\mathbb{P}\{Q \in B_c\} - \mathbb{P}\{N(0_{p+1}, I_{p+1}) \in B_c\}| = o(1)$$

uniformly in c . Now, using that $A^T A$ is block-diagonal, the asymptotic probability can be written as

$$\begin{aligned} & \mathbb{P}\{N(0_{p+1}, I_{p+1}) \in B_c\} \\ = & \mathbb{P}\left\{N\left(\begin{pmatrix} X^T \mu \\ \eta^T \mu \end{pmatrix}, \sigma^2 A^T A\right) \in E \times (-\infty, c]\right\} \\ = & \mathbb{P}\{N(X^T \mu, (1 + \gamma^2)\sigma^2 X^T X) \in E\} \mathbb{P}\{N(\eta^T \mu, (1 + \gamma^{-2})\sigma^2 \|\eta\|^2) \leq c\}. \end{aligned}$$

To conclude this part, let $c = kx + \eta^T \mu$ for an arbitrary x independent of n , where $k = (1 + \gamma^{-2})^{1/2} \sigma \|\eta\|$. We have that

$$\mathbb{P}\{k^{-1}(\eta^T V - \eta^T \mu) \leq x \mid X^T U \in E\} = \frac{\mathbb{P}(X^T U \in E, \eta^T V \leq c)}{\mathbb{P}(X^T U \in E, \eta^T V < \infty)}.$$

Using the normal approximation to the numerator and denominator, and the fact that the denominator is bounded away from zero, it follows that

$$\mathbb{P}\{k^{-1}(\eta^T V - \eta^T \mu) \leq x \mid X^T U \in E\} \rightarrow \mathbb{P}\{N(0, 1) \leq x\}.$$

This establishes convergence in distribution under a known variance assumption and a convex selection set. Extending the result to the conditions of the theorem is straightforward. First, if $\hat{\sigma} \neq \sigma$, write $V = \tilde{V} + \gamma^{-1}(\sigma - \hat{\sigma})Z$, with $\tilde{V} = Y - \gamma^{-1}\sigma Z$. Then,

$$k^{-1}(\eta^T V - \eta^T \mu) = k^{-1}(\eta^T \tilde{V} - \eta^T \mu) + k^{-1}\gamma^{-1}(\sigma - \hat{\sigma})\eta^T Z.$$

We have shown that, conditionally on selection, the first term converges in distribution to a standard normal. The second one is bounded by a constant times $(\sigma - \hat{\sigma})\|\eta\|^{-1}\eta^T Z$. Note that, unconditionally on selection, $\sigma - \hat{\sigma} = o_p(1)$, by assumption, and $\|\eta\|^{-1}\eta^T Z = O_p(1)$, as it has zero mean and unit variance. Hence, the last term is $o_p(1)$ unconditionally on selection. However, since the selection probability is bounded away from zero, it follows trivially the last term converges to zero also conditionally on selection, as

$$\mathbb{P}(|(\sigma - \hat{\sigma})\|\eta\|^{-1}\eta^T Z| > \delta \mid S = s_n) \leq \frac{\mathbb{P}(|(\sigma - \hat{\sigma})\|\eta\|^{-1}\eta^T Z| > \delta)}{\mathbb{P}(S = s_n)} \rightarrow 0$$

as $n \rightarrow \infty$ for any positive δ . The result follows from Slutsky's theorem.

Finally, suppose that $E = E_1 \cup E_2$, with E_1 and E_2 convex. We can write

$$\begin{aligned} & \mathbb{P}\{k^{-1}(\eta^T V - \eta^T \mu) \leq x \mid X^T U \in E\} \\ = & \frac{\mathbb{P}(X^T U \in E_1, \eta^T V \leq c) + \mathbb{P}(X^T U \in E_2, \eta^T V \leq c) - 2\mathbb{P}(X^T U \in E_1 \cap E_2, \eta^T V \leq c)}{\mathbb{P}(X^T U \in E_1) + \mathbb{P}(X^T U \in E_2) - 2\mathbb{P}(X^T U \in E_1 \cap E_2)} \end{aligned}$$

and use the normal approximation as before, since all the events involved are convex. When the union involves more than two sets we can proceed recursively.

Chapter 6

Numerical implementation

Implementation of the conditional approach is sometimes hindered by the need to evaluate the selection probability $\varphi(\theta)$ in a sufficiently large region of the parameter space, especially when the inferential model is multidimensional and the selection event is non-standard. In this chapter we explore the use of simulation-based methods to tackle this problem.

6.1 An Importance Sampling algorithm

Likelihood methods based on the selective distribution, including selection-adjusted Bayesian methods, require evaluating likelihood ratios $f_S(y; \theta_1)/f_S(y; \theta_2)$ at several pairs $(\theta_1, \theta_2) \in \Theta^2$ for an arbitrary $y \in \mathcal{Y}$. Assuming that the unconditional density $f(y; \theta)$ is easy to evaluate, the challenge lies on evaluation of the probability ratios

$$g(\theta_1, \theta_2) = \frac{\varphi(\theta_1)}{\varphi(\theta_2)} \tag{6.1}$$

for a potentially large set of pairs (θ_1, θ_2) . For cases where the data are normally distributed and the selection event is given by $p(y) = \mathbf{1}(y \in E)$, where E is an affine set, an analytic approximation to (6.1) was proposed by [Panigrahi et al. \(2020\)](#). This approximation enjoys some interesting theoretical properties and is fast to implement. However, it does not offer any finite-sample guarantees and its scope of applicability is limited.

Monte Carlo methods, as ubiquitous as they are in other areas, remain largely unexplored in the selective inference literature. There are, presumably, two main reasons for this. The first one is that the need to approximate a wide range of integrals makes it necessary to work with multiple proposal distributions, which need to be combined efficiently. The

second one is that the density of these distributions is usually only known up to a multiplicative constant. In this chapter we propose a simple extension of the Importance Sampling algorithm that circumvents these two issues. In the remainder of this work we will assume that the data are from a continuous distribution and that the selection event is of the form $p(y) = \mathbf{1}(y \in E)$ for some measurable $E \subseteq \mathcal{Y}$, so that

$$\varphi(\theta) = \int_E f(y; \theta) dy.$$

Suppose that $\{q_j: j = 1, \dots, m\}$ are the densities of m distributions supported on the selection event E , which can only be evaluated up to an unknown multiplicative constant. Typically they will be standard distributions, such as normals, truncated to E . We assume that a batch of n independent Monte Carlo samples $\{Y_{ij}: i = 1, \dots, n_j, j = 1, \dots, m\}$ from the proposals is available for estimation of g , with Y_{ij} coming from q_j , and $n_1 + \dots + n_m = n$.

The standard way of handling samples from several proposal distributions for integral approximation is via Multiple Importance Sampling (Veach and Guibas, 1995; Veach, 1997), which constructs a global estimator for an integral $\int_E f(y; \theta) dy$ of interest by considering weighted averages of the ratios $\{r_{ij} = f(Y_{ij}; \theta)/q_j(Y_{ij})\}$. Unfortunately, this idea cannot be applied to the problem at hand, because in our framework each r_{ij} can only be computed up to an unknown multiplicative constant depending on j , and therefore these quantities cannot be combined in a fair way.

As a natural option, we might consider approximating $g(\theta_1, \theta_2)$ by a combination of the estimators

$$\hat{g}_j(\theta_1, \theta_2) = \left\{ \sum_{i=1}^{n_j} \frac{f(Y_{ij}; \theta_1)}{q_j(Y_{ij})} \right\}^{-1} \sum_{i=1}^{n_j} \frac{f(Y_{ij}; \theta_2)}{q_j(Y_{ij})}, \quad j = 1, \dots, m. \quad (6.2)$$

These estimators are computable because the unknown normalising constants of q_j in (6.2) cancel each other out. We could, for example, take their weighted arithmetic mean with each weight proportional to an estimate of $1/\text{Var}\{\hat{g}_j(\theta_1, \theta_2)\}$. However, this approach can be inefficient, because it forces $f(y; \theta_1)$ and $f(y; \theta_2)$ to be weighted equally across the different proposals, and the optimality of each proposal might be different for each integrand. Ideally, we would like the weights to depend not only on the proposal distributions, but also on the parameter.

Although we cannot do this directly, because an uneven weighting of the summands in the numerator and denominator of (6.2) would prevent the unknown normalising constants from cancelling out, parameter-dependent weights can still be derived by considering a sequence of integrands acting as a “bridge” between $f(y; \theta_1)$ and $f(y; \theta_2)$. The idea is the

following. For a given pair (θ_1, θ_2) , let $h^{(0)}(y) = f(y; \theta_1)$, $h^{(1)}(y), \dots, h^{(Z-1)}(y)$, $h^{(Z)}(y) = f(y; \theta_2)$ be a sequence of integrable functions over E such that $h^{(z)} \approx h^{(z+1)}$ for all $z = 0, \dots, Z-1$. We can construct an estimator for $g(\theta_1, \theta_2)$ by writing it as a telescoping product,

$$g(\theta_1, \theta_2) = \frac{\int_E h^{(0)}(y) dy}{\int_E h^{(1)}(y) dy} \times \frac{\int_E h^{(1)}(y) dy}{\int_E h^{(2)}(y) dy} \times \dots \times \frac{\int_E h^{(Z-1)}(y) dy}{\int_E h^{(Z)}(y) dy}, \quad (6.3)$$

and estimating each factor in (6.3) as before. That is, for each $z = 1, \dots, Z$, we estimate the z -th factor as a combination of

$$\hat{g}_j^z(\theta_1, \theta_2) = \left\{ \sum_{i=1}^{n_j} \frac{h^{(z)}(Y_{ij})}{q_j(Y_{ij})} \right\}^{-1} \sum_{i=1}^{n_j} \frac{h^{(z-1)}(Y_{ij})}{q_j(Y_{ij})}, \quad j = 1, \dots, m. \quad (6.4)$$

The advantage of this procedure with respect to the previous one is that it adds flexibility by letting the weighting scheme be different for each factor in (6.3). Since the integrands in each factor are assumed to be similar, a good proposal distribution for the numerator will also be effective for the denominator, and vice versa.

There are many ways in which the estimators (6.4) can be combined. A simple strategy is to take, for each z , a weighted arithmetic average of $\{\log \hat{g}_j^z(\theta_1, \theta_2) : j = 1, \dots, m\}$. This is, to consider estimators of the form

$$\log \hat{g}(\theta_1, \theta_2) = \sum_{z=1}^Z \sum_{j=1}^m w_j^z \log \hat{g}_j^z(\theta_1, \theta_2),$$

where $\{w_j^z \in \mathbb{R} : j = 1, \dots, m, z = 1, \dots, Z\}$ is a set of (possibly data-dependent) weights such that $\sum_{j=1}^m w_j^z = 1$ for all z . We work on the log-scale because it allows for a simpler analysis and because working in the logarithmic scale is usually more stable from a numerical point of view.

Using intermediate functions between the densities of interest is the basic idea underlying two important classes of integral ratio estimators: path sampling (Binder, 1986; Gelman and Meng, 1994; Ogata, 1989) and bridge sampling (Meng and Wong, 1996). A problem with these methods is that they require sampling from specific proposals depending on the integrand, which in our problem may not be feasible for some choices of E . In addition, they require a new set of samples for each pair (θ_1, θ_2) , which would make them very expensive to implement.

A simple class of bridging functions can be derived by taking any smooth invertible function $t: \mathbb{R} \rightarrow \mathbb{R}$ and defining

$$h^{(z)}(y) = t^{-1} \left\{ (1 - zZ^{-1})t[f(y; \theta_1)] + zZ^{-1}t[f(y; \theta_2)] \right\},$$

or

$$h^{(z)}(y) = f(y; t^{-1}\{(1 - zZ^{-1})t(\theta_1) + zZ^{-1}t(\theta_2)\}). \quad (6.5)$$

Common choices in the Monte Carlo literature include the arithmetic path, with $t(x) = x$, and the geometric path, with $t(x) = \log(x)$. Another strategy, which was observed to perform well empirically, proceeds as follows. We define $h^{(1)}(y)$ as the density in $\{q_1(y), \dots, q_m(y), f(y; \theta_2)\}$ that is closest, in some appropriate sense, to $f(y; \theta_1)$. If this density is $f(y; \theta_2)$ we simply take $Z = 1$ and no bridging functions are used. Otherwise, the selected density is removed from the set and $h^{(2)}(y)$ is defined analogously, and this process continues until all the proposals are selected or $f(y; \theta_2)$ is reached.

As for the choice of weights, we propose the following approach. Let $w_j = (w_j^1, \dots, w_j^Z)^T$ and let V_j denote the covariance matrix of $[\log \hat{g}_j^1(\theta_1, \theta_2), \dots, \log \hat{g}_j^Z(\theta_1, \theta_2)]^T$, which we assume to be finite for all j . If the weights were deterministic, the variance of $\log \hat{g}(\theta_1, \theta_2)$ would be given by

$$\sum_{j=1}^m w_j^T V_j w_j. \quad (6.6)$$

Subject to $\sum_{j=1}^m w_j = 1_Z$, where 1_Z is a Z -vector of ones, the minimiser of (6.6) is given by

$$w_j = V_j^{-1} \left(\sum_{k=1}^m V_k^{-1} \right)^{-1} 1_Z, \quad j = 1, \dots, m. \quad (6.7)$$

This is a routine Lagrange multipliers exercise. We propose using a sample version of (6.7),

$$\hat{w}_j = \hat{V}_j^{-1} \left(\sum_{k=1}^m \hat{V}_k^{-1} \right)^{-1} 1_Z, \quad j = 1, \dots, m, \quad (6.8)$$

where $\hat{V}_1, \dots, \hat{V}_m$ are estimates of $V_1, \dots, V_m$ constructed from the Monte Carlo samples. These estimates can be derived using the Delta Method. A simple calculation shows that the approximation of the (z_1, z_2) entry of V_j provided by the Delta Method is given by

$$\frac{1}{n_j} \mathbb{E}_{q_j} \left[\left\{ \frac{h^{(z_1-1)}(Y)}{\gamma_j^{(z_1-1)} q_j(Y)} - \frac{h^{(z_1)}(Y)}{\gamma_j^{(z_1)} q_j(Y)} \right\} \times \left\{ \frac{h^{(z_2-1)}(Y)}{\gamma_j^{(z_2-1)} q_j(Y)} - \frac{h^{(z_2)}(Y)}{\gamma_j^{(z_2)} q_j(Y)} \right\} \right],$$

where

$$\gamma_j^{(z)} = \int_E h^{(z)}(y) dy,$$

which can in turn be estimated by

$$\frac{1}{n_j^2} \sum_{i=1}^{n_j} \left\{ \frac{h^{(z_1-1)}(Y_{ij})}{\hat{\gamma}_j^{(z_1-1)} q_j(Y_{ij})} - \frac{h^{(z_1)}(Y_{ij})}{\hat{\gamma}_j^{(z_1)} q_j(Y_{ij})} \right\} \times \left\{ \frac{h^{(z_2-1)}(Y_{ij})}{\hat{\gamma}_j^{(z_2-1)} q_j(Y_{ij})} - \frac{h^{(z_2)}(Y_{ij})}{\hat{\gamma}_j^{(z_2)} q_j(Y_{ij})} \right\},$$

where

$$\hat{\gamma}_j^{(z)} = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{h^{(z)}(Y_{ij})}{q_j(Y_{ij})}.$$

A drawback of (6.8) is that it can have a very large variance due to the fact that the covariance matrices tend to be small in norm, especially when Z is large. To mitigate this problem we suggest adding a small ridge term to the estimated covariances, as follows:

$$\hat{w}_j(\lambda) = (\hat{V}_j + \lambda I_Z)^{-1} \left\{ \sum_{k=1}^m (\hat{V}_k + \lambda I_Z)^{-1} \right\}^{-1} 1_Z, \quad j = 1, \dots, m,$$

where λ is a positive number and I_Z is the $Z \times Z$ identity matrix. Large values of λ will result in very stable but potentially very biased estimators of the optimal coefficients, while small ones may be too unstable. To tune the ridge parameter λ we can use the following Bootstrap algorithm: for a given size B we construct B Bootstrap replicates of $[\log \hat{g}_j^1(\theta_1, \theta_2), \dots, \log \hat{g}_j^Z(\theta_1, \theta_2)]^T$, $[\log \hat{g}_j^1(\theta_1, \theta_2)^b, \dots, \log \hat{g}_j^Z(\theta_1, \theta_2)^b]^T$, $b = 1, \dots, B$, by sampling uniformly from the original batch of Monte Carlo samples. We then set the value of λ to

$$\hat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \operatorname{Var} \left\{ \sum_{z=1}^Z \sum_{j=1}^m w_j^z(\lambda) \log \hat{g}_j^z(\theta_1, \theta_2)^b : b = 1, \dots, B \right\}$$

for a suitable grid of values $\Lambda = \{\lambda_1, \dots, \lambda_N\}$, where each $w_j^z(\lambda)$ is constructed from the original batch of samples.

6.2 Examples

Example 6.2.1. Exponential likelihood. Suppose that we have a random sample $Y_1, \dots, Y_5$ from an exponential distribution with rate parameter θ and that we want to approximate its selective likelihood for a given observation and the selection event $\bar{Y} > 1$. Using standard properties of the exponential distribution this probability can be seen to be $\varphi(\theta) = \mathbb{P}(\operatorname{Gamma}(5, \theta) > 5)$, so the selective likelihood admits a simple expression. Let us assume, however, that we cannot use this result and are forced to compute directly the 5-dimensional integral

$$\varphi(\theta) = \int_E f(\tilde{y}; \theta) d\tilde{y}, \quad (6.9)$$

where $E = \{\bar{y} > 1\}$. For the observed data point $y = (1.1, 1.1, 1.2, 1.3, 1.3)$, the selective likelihood function relative to $\theta = 1$, $L_S(\theta; y)/L_S(1; y)$, is plotted in Figure 6.1. The non-selective likelihood is also plotted for comparison. The dashed line corresponds to a Monte Carlo approximation of the likelihood ratios derived by approximating (6.9) using the algorithm described in this section. We used $n = 4 \times 10^4$ samples equally distributed among 4 proposals, which were exponential distributions with parameters 0.1, 0.5, 1, and

2, all truncated to E , $Z = 3$, and the bridge functions (6.5) with the arithmetic path $t(x) = x$. Ridge regularisation was not necessary.

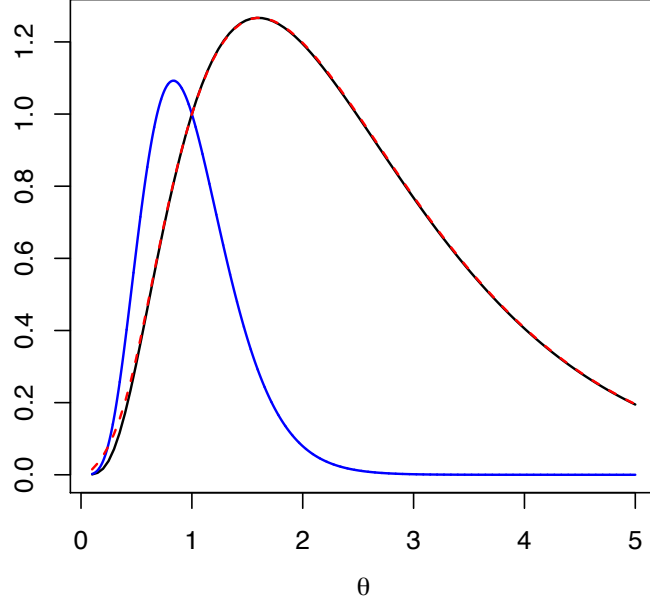


Figure 6.1: In black, $L_S(\theta; y)/L_S(1; y)$. In red, Monte Carlo approximation. In blue, non-selective version.

Example 6.2.2. *Multivariate normal posterior.* We now consider the problem of approximating the selective posterior distribution of a multivariate normal mean given an affine selection event. In order to have a tractable solution against which to compare the approximations, we will take the generative model to be $Y \sim N(\theta, I_n)$, $\theta \in \mathbb{R}^p$, and the selection event to be $E = [0, \infty)^p$. The selection probability is therefore

$$\varphi(\theta) = \prod_{i=1}^p \Phi(\theta_i).$$

Under a standard normal prior distribution $\theta \sim N(0_p, I_n)$, the selective posterior density of θ given $Y = y$ is

$$\pi(\theta|y) \propto \exp\left\{-\frac{1}{2}\|\theta\|^2\right\} \frac{\exp\left\{-\frac{1}{2}\|y - \theta\|^2\right\}}{\prod_{i=1}^p \Phi(\theta_i)}. \quad (6.10)$$

This type of posterior distributions arises in some regression settings, as described in the previous chapter. Here we will assume that interest lies in a linear contrast of the

mean, specifically $\psi = p^{-1/2} \sum_{i=1}^p \theta_i$. For $p = 10$ and $y = p^{-1/2} \mathbf{1}_p$, we approximated the marginal posterior density of ψ using a Metropolis-Hastings algorithm with normal proposal distribution in two ways. In one case we generated 10^5 posterior samples from the correct density (6.10). In the other one we run the same algorithm for 10^4 steps, and the acceptance probability at each step was estimated using the proposed numerical approximation of the probability ratios $\varphi(\theta_1)/\varphi(\theta_2)$. For this task, we generated a total of $n = 2 \times 10^4$ Monte Carlo samples, evenly distributed between $m = p$ proposal distributions: $N(-p^{-1/2} e_j, I_n)$ for $j = 1, \dots, m$, where e_j is the j -th canonical vector. The bridge functions were selected using the algorithmic approach described in the previous section, where the distance between any two distributions $N(a, I_n)$ and $N(b, I_n)$ was defined as $\|a - b\|^2$. For the Bootstrap simulations we used a grid of λ 's of 30 values between 10^{-10} and 10 evenly distributed in the log-scale, and the Bootstrap size was 50. The empirical posterior distribution functions of ψ derived from both methods can be bound in Figure 6.2. As we can see, the approximation obtained using the probability ratio approximations is very accurate.

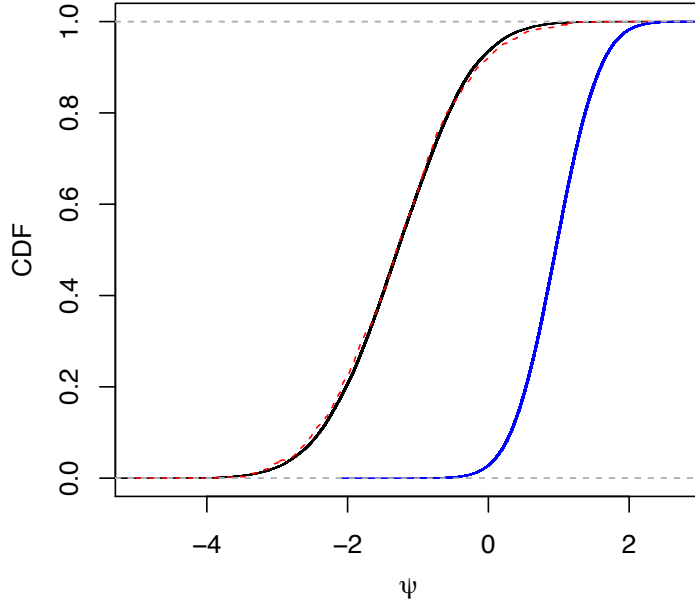


Figure 6.2: Empirical posterior distribution function of ψ obtained from Metropolis-Hastings sampling using: the correct acceptance probabilities (black), and approximations of the acceptance probabilities derived from the suggested method (red). The non-selective posterior distribution function is plotted in blue.

Chapter 7

Discussion and future work

In this thesis we have analysed different aspects of conditional selective inference. This approach represents a key strategy to inference post-selection, where the same data as used for selection of the formal statistical problem is again used for the inference itself, and has attracted considerable attention in recent years. However, several fundamental questions about conditional selective inference remain open. Our objective has been to answer some of these.

In Chapter 2, we argued that the appropriate distribution for selective inference is not determined only by the underlying generative model and the selection criterion: it needs to reflect the whole sampling process in order for the conclusions to be relevant to the observed data. We also considered the validity of the standard Bayesian approach under selection. Formally, the correct inference depends on how selection acts on the joint sampling distributions for the parameter and the data. However, determining the appropriate sampling regime for a given problem can be difficult. Furthermore, classical (unadjusted) Bayesian inference depends strongly on the assumed prior, so we argue that it is generally safer to correct for selection.

In Chapter 3, we considered two estimators for a selected normal mean. The maximum likelihood estimator is a default option in the literature when the model is one-dimensional, but it can be expensive to compute in high dimensions, as it requires solving an equation that involves the gradient of the log-selection probability. An alternative option is the maximum ordinate estimator, which is easier to implement and unbiased. However, we have seen that this can behave strangely when the selection event is not convex. In future work it would be interesting to compare these estimators with alternatives derived from non-informative priors, such as the ones proposed Chapter 4. In this chapter we also examined the asymptotic behaviour of selective models. Standard asymptotic theory

for non-selective models has a natural extension in the presence of selection. We have formalised this notion by analysing the distribution of the likelihood ratios, and have shown that under some conditions selection models can be locally approximated by selective normal models, with a selection function depending on the true parameter. The key element of this result is that ratios of selection probabilities in the original model can be uniformly approximated by ratios of normal probabilities in shrinking neighbourhoods of the true parameter. The full consequences of this are yet to be proven, but the general idea is that studying the properties of normal models, which admit a relatively simple analysis, can provide insights in more general cases. For example, we saw that, just as the uniform prior behaves poorly for inference about a selected normal mean, standard non-informative priors for other parameters in the absence of selection are similarly misbehaved when considered under selection.

In Chapter 4, we proposed two classes of non-informative prior distributions for selection models which are approximately probability matching. These priors provide some level of regularisation in low-probability regions and behave better than alternatives proposed in the literature which are independent of selection. One of the priors is data-dependent and achieves the appropriate frequentist probability matching by construction. The other is the Jeffreys prior, constructed from the selective likelihood, which provides similar results and is more computationally stable. These priors were derived heuristically and their behaviour was demonstrated empirically: further theoretical study is warranted.

In Chapter 5, we studied information-splitting methods for selective inference in the key practical context of regression models. We showed that randomisation offers an alternative to data splitting which divides the sample information more efficiently. We compared the two approaches of randomisation and data splitting with respect to their selection stability, and their selection and inferential power, showing that the dominance of the former can be substantial in settings with a limited amount of information. One limitation of randomisation, however, is that implementation requires an effective estimate of the model variance. Here we have only focused on the use of randomisation for inference after variable selection. Naturally, this idea has potential applicability as a power enhancement device in other problems where data splitting is normally used.

Finally, in Chapter 6, we have shown that simulation-based approximations of the selection probability can be successfully employed for implementation of selective inference methods in some complex scenarios. To this end, we proposed an extension of Importance Sampling that accommodates samples from several proposal distributions.

Bibliography

- Bachoc, F., Leeb, H. and Pötscher, B. M. (2017), Valid confidence intervals for post-model-selection predictors. arXiv:1412.4605v3.
- Bachoc, F., Preinerstorfer, D. and Steinberger, L. (2020), ‘Uniformly valid confidence intervals post-model-selection’, *Ann. Stat.* **48**(1), 440–463.
- Barber, R. F. and Candès, E. (2015), ‘Controlling the false discovery rate via knockoffs’, *Ann. Stat.* **43**(5), 2055–2085.
- Barber, R. F., Candès, E., Janson, L., Patterson, E. and Sesia, M. (2020), *The knockoff filter for controlled variable selection*. R package version 0.3.3.
URL: <https://CRAN.R-project.org/package=knockoff>
- Barndorff-Nielsen, O. E. (1976), ‘Noninformation’, *Biometrika* **63**(3), 567–571.
- Barndorff-Nielsen, O. E. and Cox, D. R. (1994), *Inference and Asymptotics*, Chapman and Hall/CRC, London.
- Bayarri, M. J. and Berger, J. O. (1998), ‘Robust Bayesian analysis of selection models’, *Ann. Stat.* **26**(2), 645–659.
- Bayarri, M. J. and Berger, J. O. (2000), ‘P values for composite null models’, *J. Am. Stat. Assoc.* **95**(452), 1127–1142.
- Bayarri, M. J. and DeGroot, M. H. (1987), ‘Bayesian analysis of selection models’, *J. R. Statistic. Soc. D (The Statistician)* **36**(2/3), 137–146.
- Bayati, M., Erdogdu, M. A. and Montanari, A. (2013), Estimating LASSO risk and noise level’, in C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds, ‘Advances in Neural Information Processing Systems 26 (NIPS 2013)’.
- Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013), ‘Valid post-selection inference’, *Ann. Stat.* **41**(2), 802–837.
- Bhagwat, K. V. and Subramanian, R. (1978), ‘Inequalities between means of positive operators’, *Math. Proc. Camb. Philos. Soc.* **83**(3), 393–401.

- Binder, K. (1986), *Monte Carlo Methods in Statistical Physics*, Topics in current physics, Springer, Berlin.
- Birnbaum, A. (1962), ‘On the foundations of statistical inference’, *J. Am. Stat. Assoc.* **57**, 269–326.
- Candès, E., Fan, Y., Janson, L. and Lv, J. (2018), ‘Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection’, *J. R. Statist. Soc. B* **80**(3), 551–577.
- Cohen, A. and Sackrowitz, H. B. (1989), ‘Two stage conditionally unbiased estimators of the selected mean’, *Statist. Probab. Lett.* **8**, 273–278.
- Cox, D. R. (1975), ‘A note on data-splitting for the evaluation of significance levels’, *Biometrika* **62**(2), 441–444.
- Cox, D. R. and Reid, N. (1987), ‘Parameter orthogonality and approximate conditional inference’, *J. R. Statist. Soc. B* **49**(1), 1–39.
- Datta, G. S. and Mukerjee, R. (2004), *Probability Matching Priors: Higher Order Asymptotics*, Springer, New York.
- Dawid, A. P. (1994), ‘Selection paradoxes of Bayesian inference’, in T. W. Anderson, K. T. Fang and I. Olkin, eds, ‘Multivariate analysis and its applications’, Vol. 24, Institute of Mathematical Statistics, Monograph Series., pp. 211–220.
- DiCiccio, C. J., DiCiccio, T. J. and Romano, J. P. (2020), ‘Exact tests via multiple data splitting’, *Stat. Probab. Lett.* **166**.
- Fan, J., Guo, S. and Hao, N. (2012), ‘Variance estimation using refitted cross-validation in ultrahigh dimensional regression’, *J. R. Statist. Soc. B* **74**(1), 37–65.
- Fisher, R. A. (1925), ‘Theory of statistical estimation’, *Proc. Camb. Phil. Soc.* **22**(5), 700–725.
- Fisher, R. A. (1930), ‘Inverse probability’, *Proc. Camb. Philos. Soc.* **26**, 528–535.
- Fisher, R. A. (1934a), ‘The effect of methods of ascertainment upon the estimation of frequencies’, *Ann. Eugen.* **6**, 13–25.
- Fisher, R. A. (1934b), ‘Two new properties of mathematical likelihood’, *Proc. R. Soc. Lond. A* **144**(852), 285–307.
- Fithian, W., Sun, D. L. and Taylor, J. E. (2017), ‘Optimal inference after model selection’. arXiv:1410.2597v4.
- Fraser, D. A. S., Reid, N., Marras, E. and Yi, G. Y. (2010), ‘Default priors for Bayesian and frequentist inference’, *J. R. Statistic. Soc. B* **72**(5), 631–654.

- Gelman, A. and Meng, X.-L. (1994), Path sampling for computing normalizing constants: identities and theory, Technical Report 376, Department of statistics, University of Chicago.
- Godambe, V. P. (1976), ‘Conditional likelihood and unconditional optimum estimating equations’, *Biometrika* **63**(2), 277–284.
- Godambe, V. P. (1980), ‘On sufficiency and ancillarity in the presence of a nuisance parameter’, *Biometrika* **67**(1), 155–162.
- Harville, D. A. (2021), ‘Bayesian inference is unaffected by selection: fact or fiction?’, *Am. Stat.* **00**(0), 1–7.
- Hofner, B. and Hothorn, T. (2017), *Stability selection with error control*. R package version 0.6-3.
URL: <https://CRAN.R-project.org/package=stabs>
- Hong, L., Kuffner, T. A. and Martin, R. (2018), ‘On overfitting and post-selection uncertainty assessments’, *Biometrika* **105**(1), 221–224.
- Höglund, T. (1974), ‘The exact estimate - a method of statistical estimation’, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **29**, 257–271.
- Ignatiadis, N., Klaus, B., Zaugg, J. and Huber, W. (2016), ‘Data-driven hypothesis weighting increases detection power in genome-scale multiple testing’, *Nat. Methods* **13**(7), 577–580.
- Jørgensen, B. (1994), ‘The rules of conditional inference: is there a universal definition of non-information?’, *J. Ital. Statistic. Soc.* **3**, 355–384.
- Kivaranovic, D. and Leeb, H. (2020), ‘On the length of post-model-selection confidence intervals conditional on polyhedral constraints’, *JASA* **116**(534), 1–13.
- Kozakiewicz, W. (1947), ‘On the convergence of sequences of moment generating functions’, *Ann. Math. Stat.* **18**(1), 61–69.
- Kuffner, T. A. and Young, G. A. (2018), ‘Principled statistical inference in data science’, in N. Adams, E. Cohen and Y.-K. Guo, eds, ‘Statistical Data Science’, World Scientific Publishing, pp. 21–36.
- Le Cam, L. (1960), ‘Locally asymptotically normal families of distributions’, *University of California Publications in Statistics* **3**, 37–98.
- Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016), ‘Exact post-selection inference, with application to the lasso’, *Ann. Stat.* **44**(3), 907–927.

- Lee, J. D. and Taylor, J. E. (2014), ‘Exact post model selection inference for marginal screening’, *Adv. Neural Inf. Process Syst.* **1**, 136–144.
- Lockhart, R., Taylor, J. E., Tibshirani, R. and Tibshirani, R. (2014), ‘A significance test for the lasso’, *Ann. Stat.* **42**(2), 413–468.
- Loftus, J. R. and Taylor, J. E. (2014), ‘A significance test for forward stepwise model selection’. arXiv:1405.3920v1.
- Mailhot, L. (1985), ‘Quelques aspects statistiques des lois de probabilité réelles tronquées à droite’, *Mathématiques et sciences humaines* **90**, 45–80.
- Mandel, M. and Rinott, Y. (2007), On statistical inference under selection bias, Technical report, Center for Rationality and Interactive Decision Theory, Hebrew University, Jerusalem.
- Mandel, M. and Rinott, Y. (2009), ‘A selection bias conflict and frequentist versus bayesian viewpoints’, *Am. Stat.* **63**(3), 211–217.
- Markovic, J., Taylor, J. and Taylor, J. (2019), Inference after black box selection. arXiv:1901.09973v1.
- Meinshausen, N. and Bühlmann, P. (2010), ‘Stability selection’, *J. R. Statist. Soc. B* **72**(4), 417–473.
- Meng, X.-L. and Wong, W. H. (1996), ‘Simulating ratios of normalizing constants via a simple identity’, *Statist. Sinica* **6**, 831–860.
- Ogata, Y. (1989), ‘A Monte Carlo method for high dimensional integration’, *Numer. Math.* **55**, 137–157.
- Owen, D. B. (1980), ‘A table of normal integrals’, *Commun. Statist. - Simula. Computa.* **9**(4), 389–419.
- Panigrahi, S. and Taylor, J. (2018), ‘Scalable methods for Bayesian selective inference’, *Electron. J. Stat.* **12**, 2355–2400.
- Panigrahi, S., Taylor, J. and Weinstein, A. (2020), ‘Integrative methods for post-selection inference under convex constraints’. arXiv:1605.08824v7.
- Raič, M. (2019), ‘A multivariate Berry–Esseen theorem with explicit constants’, *Bernoulli* **25**(4A), 2824–2853.
- Rao, C. R. (1985), *Weighted distributions arising out of methods of ascertainment: What population does a sample represent?*, Springer, New York, NY, pp. 543–569.

- Reid, S., Tibshirani, R. and Friedman, J. (2016), ‘A study of error variance estimation in lasso regression’, *Stat. Sin.* **26**, 35–67.
- Reitermanová, Z. (2010), ‘Data splitting’, *WDS’10 Proceedings of Contributed Papers* **1**, 31–36.
- Rinaldo, A., Wasserman, L. and G’Sell, M. (2019), ‘Bootstrapping and sample splitting for high-dimensional, assumption-lean inference’, *Ann. Stat.* **47**(6), 3438–3469.
- Rosenthal, R. (1979), ‘The file drawer problem and tolerance for null results’, *Psychol Bull.* **86**(3), 638–641.
- Rubin, D., Dudoit, S. and van der Laan, M. (2006), ‘A method to increase the power of multiple testing procedures through sample splitting’, *Stat. Appl. Genet. Mol. Biol.* **5**(19).
- Shah, R. and Samworth, R. (2013), ‘Variable selection with error control: another look at stability selection’, *J. R. Statist. Soc. B* **75**(1), 55–80.
- Shen, L. (2001), ‘An improved method of evaluating drug effect in a multiple dose clinical trial’, *Statist. Medicine* **20**, 1913–1929.
- Snee, R. D. (1977), ‘Validation of regression models: methods and examples’, *Technometrics* **19**, 415–428.
- Stallard, N. and Todd, S. (2005), ‘Point estimates and confidence regions for sequential trials involving selection’, *J. Stat. Plan. Inference* **135**, 402–419.
- Stallard, N., Todd, S. and Whitehead, J. (2008), ‘Estimation following selection of the largest of two normal means’, *J. Stat. Plan. Inference* **138**, 1629–1638.
- Tian, X. (2020), ‘Prediction error after model search’, *Ann. Stat.* **48**(2), 763–784.
- Tian, X. and Taylor, J. E. (2018), ‘Selective inference with a randomized response’, *Ann. Stat.* **46**(2), 679–710.
- Tibshirani, R., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2018), ‘Uniform asymptotic inference and the bootstrap after model selection’, *Ann. Stat.* **46**(3), 1255–1287.
- Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S. and Markovic, J. (2019), *selectiveInference: Tools for post-selection inference*. R package version 1.2.5.
URL: <https://CRAN.R-project.org/package=selectiveInference>
- van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.
- Veach, E. (1997), Robust Monte Carlo methods for light transport simulation, PhD thesis, Stanford University.

- Veach, E. and Guibas, L. J. (1995), ‘Optimally combining sampling techniques for Monte Carlo rendering’, in ‘Proceedings of the 22nd annual conference on computer graphics and interactive techniques’, pp. 419–428.
- Wasserman, L. and Roeder, K. (2009), ‘High-dimensional variable selection’, *Ann. Stat.* **37**(5A), 2178–2201.
- Xie, M. and Singh, K. (2013), ‘Confidence distribution, the frequentist distribution estimator of a parameter: a review’, *Int. Stat. Rev.* **81**(1), 3–39.
- Yekutieli, D. (2012), ‘Adjusted Bayesian inference for selected parameters’, *J. R. Statist. Soc. B* **74**(3), 515–541.
- Young, G. A. and Smith, R. L. (2005), *Essentials of Statistical Inference*, Cambridge University Press, United Kingdom.